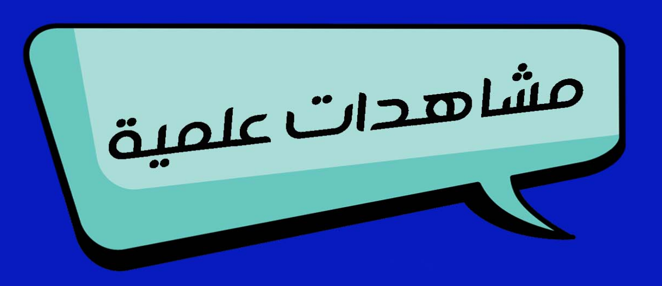


البيانات الضخمة

بقلم: دون إي هولمز

إعداد وتحرير: رأفت علام

مكتبة المشرق الإلكترونية





مشاهدات علمية
البيانات الضخمة
بقلم: دون إي هولمز
إعداد وتحرير: رأفت علام
مكتبة المشرق الإلكترونية

البيانات الضخمة

أفكار: دون إي هولمز

ترجمة فريق الترجمة بمكتبة المشرق الإلكترونية
إعداد وتحرير: رأفت علام
مكتبة المشرق الإلكترونية

صدر في يناير 2024 عن مكتبة المشرق الإلكترونية – مصر

شكر وتقدير

عندما قلتُ لبيتر إني أودُّ أن أشكره على مساهمته في هذا الكتاب، اقترح عليَّ الآتي: «أودُّ أن أشكر بيتر هاربر، الذي لولا استخدامه المتفاني للمدقِّق الإملائي، لكان هذا كتابًا مختلفًا». كما أودُّ أن أشكره على خبرته في إعداد القهوة، وما يتمتَّع به من حس الدعابة! هذا الدعم، في حد ذاته، لا يُقدَّر بثمن، ولكن، ما فعله بيتر يفوق ذلك بكثير، ولن أُبالغ حين أقول إنه لولا تحفيزه المستمر ومساهماته البنَّاءة، لم يكن لهذا الكتاب أن يرى النور.

دون هولمز
أبريل ٢٠١٧

تمهيد

تندرج الكتب التي تتناول موضوع البيانات الضخمة ضمن أحد تصنيفَين: إمَّا أنها لا تُقدِّم أيَّ تفسيراتٍ عن آلية عمل البيانات الضخمة، وإمَّا أنها تكون كتبًا دراسية مُتخصِّصة في مجال الرياضيات لا تصلح إلا لطلاب الدراسات العليا. يهدف هذا الكتاب إلى تقديم بديل عن طريق توفير مقدمة إلى آلية عمل البيانات الضخمة وكيفية تغييرها للعالم من حولنا، وتأثيرها في حياتنا اليومية، وفي عالم الأعمال.

كانت البيانات تعني في الماضي المستندات والأوراق، وربما بعض الصور، ولكنها أصبحت تعني الآن أكثر من ذلك بكثير. تُنتج مواقع شبكات التواصل الاجتماعي كلَّ دقيقة كمياتٍ كبيرةً من البيانات على هيئة صور، ومقاطع فيديو، وأفلام. ويُنتج التسوُّق عبر الإنترنت بياناتٍ عندما نُدخل عناويننا وبيانات بطاقاتنا الائتمانية. ووصلنا حاليًّا إلى مرحلة أصبح فيها جمع البيانات وتخزينها يتطوَّر على نحو لم نكن نتخيَّله منذ بضعة عقود مضت، ولكن، كما سنرى في هذا الكتاب، فإنَّ أساليب تحليل البيانات الجديدة تُحوِّل هذه البيانات إلى معلومات مفيدة. أثناء تأليف هذا الكتاب، تبيَّن لي أنه لا يمكن مناقشة موضوع البيانات الضخمة على نحوٍ مجدٍ من دون التطرق مرارًا وتكرارًا إلى عمليات جمعها، وتخزينها، وتحليلها، واستخدامها من قِبَل الشركات التجارية الكبرى. وبما أن الأقسام البحثية في شركات على غرار جوجل وأمازون هي المنوط بها مسئولية الكثير من التطورات الرئيسية في مجال البيانات الضخمة، فسوف نذكرها مرارًا وتكرارًا.

يُعرِّف الفصل الأول القارئ بتنوُّع البيانات بوجه عام، قبل أن يشرح كيف أدَّى العصر الرقمي إلى تغييرات في طريقة تعريفنا للبيانات. تُطرَح البيانات الضخمة على نحو غير رسمي عبر فكرة انفجار البيانات، والتي تتضمَّن علوم الكمبيوتر، وعلم الإحصاء، ونقاط الالتقاء بينهما. في الفصول من الثاني إلى الرابع، استخدمتُ الأشكال التخطيطية على نحو مكثَّف لمساعدتي في شرح بعض من الأساليب الجديدة التي تتطلَّبها البيانات الضخمة. ويتحدَّث الفصل الثاني عن أسباب تميُّز البيانات الضخمة، وهو ما يقودنا إلى تعريفٍ أكثر تحديدًا لها. وفي الفصل الثالث، نُناقش المشكلات المتعلِّقة بتخزين البيانات الضخمة وإدارتها. يُدرك أغلب الناس الحاجة إلى الاحتفاظ بنسخة احتياطية من البيانات على أجهزة الكمبيوتر الشخصية. ولكن، كيف نفعل ذلك مع الكميات الهائلة من البيانات التي يجري إنتاجها حاليًّا؟ للإجابة عن هذا السؤال، سنتناول تخزين قواعد البيانات وفكرة توزيع المهام على مجموعات مترابطة من أجهزة الكمبيوتر. يبرهن الفصل الرابع على أن البيانات الضخمة لا تكون مفيدةً إلا إذا تمكَّنا من استخراج معلوماتٍ مفيدة منها. ونعطي لمحةً عن كيفية تحوُّل البيانات إلى معلومات باستخدام شروح مبسَّطة للعديد من الأساليب الراسخة.

بعد ذلك ننتقل إلى مناقشةٍ أكثر تفصيلًا عن تطبيقات البيانات الضخمة؛ حيث نبدأ في الفصل الخامس بدَور البيانات الضخمة في مجال الطب. ويحلِّل الفصل السادس الممارسات التجارية باستخدام دراستَي حالة عن شركتَي أمازون ونتفليكس، تُبرز كل منهما سمات مختلفة للتسويق باستخدام

البيانات الضخمة. يتناول الفصل السابع بعض مشكلات الأمان التي تحيط بالبيانات الضخمة وأهمية التشفير. أصبحت سرقة البيانات مشكلةً كبيرة، وسنتناول بعض القضايا التي تناولتها الصحف، بما فيها قضية سنودن وويكيليكس. ويُختَتم الفصلٍ بتوضيح كيف أن جرائم الإنترنت أضحت من المشكلات التي يتوجَّب على البيانات الضخمة حَلُّها. في الفصل الثامن والأخير، سنتناول كيف تُغيِّر البيانات الضخمة المجتمع الذي نعيش فيه؛ وذلك من خلال إنشاء الروبوتات المتطوِّرة ودورها في مكان العمل. ونختتم الكتاب بتناول المنازل الذكية والمدن الذكية المستقبلية.

لا يمكن أن نستوفي في مقدمة قصيرة جدًّا كل شيءٍ في هذا الصدد، ومن ثم، آمل أن يواصل القارئ مطالعة الموضوعات التي تهمه من خلال الاستعانة بالتوصيات التي أوردناها في جزء «قراءات إضافية».

الفصل الأول

انفجار البيانات

ما البيانات؟

في عام ٤٣١ قبل الميلاد، أعلنت أسبرطة الحرب على أثينا. يصف ثيوسيديدز، في روايته عن الحرب، كيف خطَّطت القوات البلاتية المُحاصَرة الموالية لأثينا للهرب عن طريق تسلُّق الجدار المحيط ببلاتايا الذي بنته القوات البيلوبونيسية تحت القيادة الأسبرطية. ولكي يتمكَّنوا من ذلك، كانوا يحتاجون إلى معرفة ارتفاع الجدار حتى يصنعوا سلالم ذات طول مناسب. كانت أجزاء كثيرة من الجدار البيلوبونيسي مغطاةً بالجص الخشن، إلا أنهم عثروا على جزء منه حيث كان الطوب لا يزال ظاهرًا بوضوح، وكُلِّف عدد كبير من الجنود بمهمة عدِّ طبقات هذا الطوب المكشوف. كان العمل بمنأًى آمن عن هجمات العدو، يقتضي حتمًا وجود أخطاء، ولكن، كما يُوضِّح ثيوسيديدز، مع التسليم بإجراء العد مراتٍ عديدة، فإن النتيجة الأكثر تكرارًا ستكون هي الصحيحة. هذا العدد الأكثر تكرارًا، والذي سنطلق عليه الآن «المنوال»، استُخدِم بعد ذلك لحساب ارتفاع الجدار؛ فقد كان البيلوبونيسيون يعرفون حجم الطوب المحلي المُستخدم، وصُنعت السلالم ذات الارتفاع المطلوب لتسلُّق الجدار. ومكَّن هذا قوةً مؤلَّفة من عدة مئات من الرجال من الهرب، ويمكن اعتبار هذه الحادثة أكثر مثال لافت للنظر في تاريخ جمع البيانات وتحليلها. ولكن، يرجع جمع البيانات، وتخزينها، وتحليلها إلى ما قبل عصر ثيوسيديدز بقرون، كما سنرى لاحقًا.

وُجدت علامات محفورة على عصي، وأحجار، وعظام، تعود إلى العصر الحجري القديم الأعلى. ويُعتقد أن هذه الحزوز كانت بغرض تمثيل البيانات المُخزَّنة كعلامات إحصاء، ولكن، لا يزال هذا الاعتقاد مفتوحًا للنقاش الأكاديمي. ولعل أشهر مثال على ذلك هو عَظْمة إشانجو، التي عُثر عليها في جمهورية الكونغو الديمقراطية عام ١٩٥٠، ويُقدَّر عمرها بحوالي ٢٠ ألف سنة. تعدَّدت التفسيرات لهذه العظمة المحزَّزة ما بين كونها آلةً حاسبة أو رزنامة، في حين فضَّل آخرون تفسير وجود العلامات عليها بأنها بغرض إحكام مسكها. عظمة ليبومبو، المُكتشفة في سبعينيات القرن العشرين في سوازيلاند، أقدم من سابقتها؛ حيث يرجع تاريخها إلى حوالي ٣٥ ألف سنة قبل الميلاد. تحتوي هذه الشظية من عظمة قرد البابون، على تسعة وعشرين خطًّا عرضيًّا تشبه كثيرًا عصي التقويم التي لا يزال شعب البوشمن يستخدمونها في أقاصي ناميبيا، ما يدل على أنها ربما كانت حقًّا طريقةً مُستخدمة لمتابعة البيانات التي تهم حضارتهم.

على الرغم من أن تفسير هذه العظام المحزَّزة لا يزال مفتوحًا للتخمين، فإننا نعلم أن أحد أول استخدامات البيانات الجيدة التوثيق كان الإحصاء السكاني الذي أجراه البابليون عام ٣٨٠٠ قبل الميلاد. وثَّق هذا الإحصاء السكاني بطريقة منهجية عددَ السكان والسلع، مثل الحليب والعسل؛ من أجل توفير المعلومات اللازمة لحساب الضرائب. استخدم المصريون القدماء أيضًا البيانات، في

صورة كتابات هيروغليفية على الخشب أو ورق البردي؛ من أجل تسجيل تسليم البضائع ومتابعة الضرائب. ولكن، الأمثلة الأولى على استخدام البيانات ليست قاصرة، بأي حال من الأحوال، على أوروبا وأفريقيا. كان شعب الإنكا، ومَن سبقهم من شعوب أمريكا الجنوبية، حريصين على تسجيل الإحصاءات لأغراض ضريبية وتجارية، واستخدموا نظامًا دقيقًا ومعقدًا من الخيوط المعقودة الملوَّنة، كانت تُسمى «كيبو»؛ ليكون بمثابة نظام محاسبة عشري. ترجع هذه الخيوط المعقودة، المنسوجة من وبر الإبل أو القطن المصبوغ بألوان فاتحة، إلى الألفية الثالثة قبل الميلاد، ومع أنه من المعروف أن إجمالي ما نجا من الغزو الإسباني وما تلاه من محاولات لطمس هذه الخيوط يقل عن ألف خيط، فإنها تُعد من أوائل الأمثلة المعروفة على أنظمة تخزين البيانات العملاقة. يجري حاليًّا تطوير خوارزميات الكمبيوتر في محاولة لتفسير المعنى الكامل لخيوط «الكيبو»، وتعزيز فهمنا لكيفية استخدامها قديمًا.

على الرغم من إمكانية التفكير في هذه الأنظمة المبكرة ووصفها بأنها تستخدم البيانات، فإن كلمة Data (أي بيانات) هي في الحقيقة صيغة جمع ذات أصل لاتيني، ومفردها Datum. ونادرًا ما تُستخدم كلمة Datum في العصر الحالي؛ ومن ثمَّ تُستخدم كلمة Data تعبيرًا عن صيغتَي المفرد والجمع. ينسب «قاموس أكسفورد الإنجليزي» أول استخدام معروف للكلمة إلى الكاهن الإنجليزي هنري هاموند خلال القرن السابع عشر، وكان ذلك في منشور ديني مثير للجدل نُشر عام ١٦٤٨. استخدم هاموند في هذا المنشور عبارة «كومة من البيانات»، بمفهوم لاهوتي، في إشارة إلى الحقائق الدينية التي لا تقبل الجدل. ولكن، على الرغم من أن هذا المنشور يبرز بوصفه أنه يمثل أول استخدام لكلمة «بيانات» في اللغة الإنجليزية، فإنه لا يتضمَّن استخدامها بالمفهوم العصري الذي يعني الحقائق والأرقام المتعلِّقة بمجموعة معينة هي موضع اهتمام. تعود نشأة مصطلح «البيانات»، بمفهومه الحالي، إلى الثورة العلمية في القرن الثامن عشر بقيادة عمالقة المفكِّرين أمثال بريستلي، ونيوتن، ولافوازييه، وبحلول عام ١٨٠٩، بعد أعمال علماء الرياضيات الأوائل، أرسى كلٌّ من جاوس ولابلاس أسسًا رياضية للغاية للمنهجية الإحصائية الحديثة.

على مستوًى أكثر عملية، جُمعت كمية هائلة من البيانات خلال تفشي وباء الكوليرا عام ١٨٥٤ في شارع برود بمدينة لندن، ما مكَّن الطبيب جون سنو من إعداد مخطط بياني عن حالة التفشي هذه. وبذلك، تمكَّن من دعم فرضيته أن الماء الملوَّث تسبَّب في انتشار المرض، وإثبات أن المرض لا ينتقل عبر الهواء كما كان يُعتقد سابقًا. بجمع البيانات من السكان المحليين، أثبت أن المصابين بالمرض كانوا يستخدمون جميعهم مضخة المياه العمومية نفسها؛ ومن ثمَّ أقنع المسئولين المحليين عن الأبرشية بإغلاقها، المهمة التي أنجزوها عن طريق إزالة مقبض المضخة. بعد ذلك، وضع سنو خريطة، صارت مشهورةً حاليًّا، تُظهر أن المرض ظهر في مجموعات عنقودية مترابطة تحيط بمضخة برود ستريت. واصلَ سنو العمل في هذا الصدد، حيث راح يجمع البيانات ويحلِّلها، واشتهر بكونه أحد اختصاصيي الأوبئة الرواد.

بعد البحث الذي قدَّمه جون سنو، تزايد استخدام اختصاصيي الأوبئة وعلماء الاجتماع للبيانات الديموجرافية اللازمة للأغراض البحثية، وأثبت الإحصاء السكاني الذي أصبح يُجرى الآن في الكثير من الدول أنه مصدر مفيد لهذه المعلومات. على سبيل المثال، تُجمَع الآن كل البيانات الخاصة بمعدلات المواليد والوَفَيَات، وتكرار الإصابة بمختلِف الأمراض، وإحصاءات الدخل والجريمة، ولم

يكن الحال على هذا المنوال قبل القرن التاسع عشر. أصبح الإحصاء السكاني، الذي يُجرى كل عشرة أعوام في أغلب الدول، يجمع كميات متزايدةً من البيانات، وهو ما زاد، في نهاية المطاف، عمَّا يمكن تسجيله باليد أو بأجهزة الإحصاء البسيطة التي كانت مستخدمةً سابقًا. تمَّ التصدي جزئيًّا لتحدي معالجة هذه الكميات المتزايدة من بيانات الإحصاء السكاني من قِبل هيرمان هوليريث أثناء عمله في مكتب تعداد الولايات المتحدة.

بحلول موعد التعداد الأمريكي لعام ١٨٧٠، أصبحت أجهزة إحصاء بسيطة قيد الاستخدام، إلا أن هذا لم يُحقِّق إلا نجاحًا محدودًا في تقليل كم العمل الذي يؤديه مكتب التعداد. ولكن حدثت طفرة جاءت في أوانها قبل تعداد عام ١٨٩٠ عندما استُخدِمَت آلة تبويب البطاقات المُثقَّبة التي اخترعها هيرمان هوليريث لتصنيف البيانات ومعالجتها. كانت معالجة بيانات التعداد الأمريكي تستغرق في المعتاد ثماني سنوات، ولكن، باستخدام هذا الاختراع الجديد تقلَّصت هذه الفترة إلى سنة واحدة. وهكذا، أحدثت آلة هوليريث ثورةً في تحليل بيانات التعداد السكاني في جميع بلدان العالم، بما في ذلك ألمانيا، وروسيا، والنرويج، وكوبا.

بعد ذلك، باع هوليريث آلته إلى الشركة التي تطوَّرت فيما بعدُ لتصبح شركة آي بي إم، والتي طوَّرت فيما بعدُ وأنتجت سلسلةً واسعة الانتشار من آلات البطاقات المثقبة. عام ١٩٦٩، عيَّن المعهد الأمريكي للمعايير الوطنية كود هوليريث للبطاقات المثقبة (أو كود بطاقات هوليريث) بوصفه معيارًا؛ تكريمًا لهوليريث على ابتكاراته السابقة لأوانها في مجال البطاقات المثقبة.

البيانات في العصر الرقمي

قبل استخدام أجهزة الكمبيوتر على نطاق واسع، كانت بيانات التعداد السكاني، أو التجارب العلمية، أو استطلاعات رأي واستبيانات العينات المُصمَّمة بعناية تُسجَّل على الورق، العملية التي كانت تستهلك الكثير من الوقت والمال. لم يكن جمع البيانات يبدأ إلا بعدما يُقرِّر الباحثون الأسئلة التي يريدون أن تجيب عنها تجاربهم أو استطلاعاتهم؛ ومن ثمَّ، يُمكن التعامل بسهولة مع البيانات الناتجة المهيكلة للغاية، المدوَّنة على الورق في صفوف وأعمدة مرتَّبة، باستخدام طُرق التحليل الإحصائي التقليدية. بحلول النصف الأول من القرن العشرين، خُزنت بعض البيانات على أجهزة الكمبيوتر؛ ما ساعد في تخفيف جزء من هذا العمل الذي يتطلَّب الكثير من الأيدي العاملة، ولكن، بإطلاق شبكة الإنترنت العالمية (الويب) عام ١٩٨٩، وتطوُّرها السريع، زادت إمكانية إنتاج، وجمع، وتخزين، وتحليل البيانات إلكترونيًّا. بعد ذلك، ظهرت الحاجة إلى علاج المشكلات الحتمية التي نتجت عن الكم الهائل من البيانات التي أصبح من السهل الوصول إليها بفضل شبكة الويب، وسنتناول أولًا كيفية التمييز بين أنواع البيانات المختلفة.

يمكن تصنيف البيانات التي نستخرجها من شبكة الويب إلى بيانات هيكلية، أو غير هيكلية، أو شبه هيكلية.

أصبحت حاليًّا البيانات الهيكلية، من النوع المكتوب يدويًّا والمحفوظ في دفاتر أو في خزانات الملفات، تُخزَّن إلكترونيًّا في جداول بيانات أو قواعد بيانات، وتتكوَّن من جداول منسقة على هيئة جداول بيانات تتضمَّن صفوفًا وأعمدة، كل صف يمثِّل سِجلًّا، وكل عمود يمثِّل حقلًا محدَّدًا (مثل الاسم، أو العنوان، أو السن). نحن نُسهم في مخازن البيانات الهيكلية هذه عندما نُدخِل، على سبيل المثال، المعلومات الضرورية لطلب سلعة ما عبر الإنترنت. إن البيانات الهيكلية والمجدولة بعناية من السهل نسبيًّا إدارتها، وتكون قابلةً للتحليل الإحصائي؛ ذلك أنه حتى وقت قريب لم يكن من الممكن تطبيق أساليب التحليل الإحصائي إلا على البيانات الهيكلية.

على النقيض من ذلك، البيانات غير الهيكلية ليس من السهل تصنيفها، وتحتوي على صور، ومقاطع فيديو، وتغريدات، ومستندات معالجة نصوص. بمجرد انتشار استخدام شبكة الإنترنت العالمية، تبيَّن أن عددًا كبيرًا من مصادر المعلومات المحتملة ظل الوصول إليها متعذِّرًا؛ لأنها افتقدت الهيكلة المطلوبة لتطبيق أساليب التحليل القائمة. ولكن، من خلال تحديد السمات الرئيسية، يتضح أن البيانات التي تبدو للوهلة الأولى غير هيكلية قد لا تكون من دون هيكلة على الإطلاق. تحتوي رسائل البريد الإلكتروني، على سبيل المثال، على «بيانات تعريف» هيكلية في العنوان الرئيسي، ولكن الرسالة الفعلية غير الهيكلية توجد في نص الرسالة؛ ومن ثَمَّ يمكن تصنيفها على أنها بيانات شبه هيكلية. يمكن استخدام علامات بيانات التعريف، وهي في الأساس إشارات وصفية، لإضافة بعض الهيكلة إلى البيانات غير الهيكلية. إن إضافة كلمة وصفية إلى صورة على موقع إلكتروني تجعلها قابلةً للتحديد، وتُسهِّل كثيرًا من البحث عنها. توجد البيانات شبه الهيكلية أيضًا في مواقع شبكات التواصل الاجتماعي التي تستخدم الوسوم حتى يمكن تحديد الرسائل (التي هي بيانات غير هيكلية) عن موضوع مُعيَّن. إن التعامل مع البيانات غير الهيكلية أمر صعب؛ بما أنه لا يمكن تخزينها في قواعد أو جداول البيانات التقليدية، فلا بد من تطوير أدوات خاصة لاستخراج معلومات مفيدة منها. في الفصول الآتية، سنتناول كيفية تخزين البيانات غير الهيكلية.

يشير مصطلح «انفجار البيانات»، عنوان هذا الفصل، إلى الكم الهائل المتزايد من البيانات الهيكلية، وغير الهيكلية، وشبه الهيكلية التي تُنتَج كل دقيقة، وسنتناول لاحقًا بعضًا من المصادر الكثيرة المختلفة التي تُنتِج كل هذه البيانات.

مقدمة إلى البيانات الضخمة

أثناء بحثي عن المادة التي سأستخدمها في هذا الكتاب، غُمرتُ بالكمِّ غير المحدود من البيانات المتوافرة على شبكة الإنترنت — من المواقع الإلكترونية، والمجلات العلمية، والكتب الدراسية الإلكترونية. طبقًا لدراسة عالمية حديثة أجرتها شركة آي بي إم، حوالي ٢٫٥ إكسابايت من البيانات تُنتَج كل يوم. الإكسابايت الواحد يساوي 10^{18} (واحدًا متبوعًا بثمانية عشر صفرًا) بايت (أو مليون تيرابايت؛ انظر جدول الحجم بالبايت في نهاية هذا الكتاب). إذا اشتريتَ كمبيوترًا محمولًا جيدًا في وقت تأليف هذا الكتاب، فإنه سيحتوي عادةً على قرص صلب سعته التخزينية واحد أو اثنان

تيرابايت. في البداية، أشار مصطلح «البيانات الضخمة» إلى الكميات الكبيرة للغاية من البيانات التي تُنتَج في العصر الرقمي. وتشمل تلك الكميات الهائلة من البيانات، سواءٌ كانت هيكليةً أو غير هيكلية، جميع بيانات شبكة الإنترنت الناتجة عن رسائل البريد الإلكتروني، والمواقع الإلكترونية، ومواقع شبكات التواصل الاجتماعي.

حوالي ٨٠ بالمائة من بيانات العالم عبارة عن بيانات غير هيكلية في هيئة نصوص وصور؛ ومن ثَمَّ، فإنه لا يمكن التعامل معها باستخدام أساليب تحليل البيانات الهيكلية التقليدية عليها. لم يعد مصطلح «البيانات الضخمة» يُستخدم حاليًّا للإشارة إلى إجمالي كمية البيانات الناتجة والمخزَّنة إلكترونيًّا فحسب، بل أصبح يشير أيضًا إلى مجموعات البيانات الكبيرة من حيث الحجم والتعقيد، والتي تتطلَّب أساليب خوارزميةً جديدة لاستخراج معلومات مفيدة منها. تأتي مجموعات البيانات الكبيرة هذه من مصادر مختلفة؛ ولذا دعونا نتناول بعضها بمزيد من التفصيل، وكذلك البيانات التي تُنتجها.

بيانات محرِّكات البحث

عام ٢٠١٥، كان جوجل محرك البحث الأشهر على الإطلاق في جميع أنحاء العالم، وحلَّ محرك بحث بينج التابع لشركة مايكروسوفت ومحرك بحث ياهو سيرش في المركزَين الثاني والثالث، على الترتيب. عام ٢٠١٢، أحدثُ عام كانت فيه البيانات متاحةً للجمهور، بلغ حجم عمليات البحث التي تُجرى على محرك بحث جوجل وَحده ما يزيد عن ٣٫٥ مليار عملية بحث يوميًّا.

يترتَّب على إدخال كلمة أساسية ما في محرك البحث عرض قائمة بالمواقع الإلكترونية الأكثر صلة، ولكن، في الوقت نفسه، تُجمَع كمية كبيرة من البيانات. يُنتج التعقب على شبكة الويب بياناتٍ ضخمة. وكتدريب على ذلك، بحثتُ عن «سلالة كلاب بوردر كولي»، ونقرت على الموقع الإلكتروني الأول في نتائج البحث. وباستخدام أحد برامج التعقب البسيطة، وجدتُ أنه جرى إنشاء روابط إلى حوالي ٦٧ موقعًا آخر بمجرد النقر على هذا الموقع الإلكتروني. ومن أجل تعقب اهتمامات الأشخاص الذين تصفَّحوا هذا الموقع، تجري مشاركة المعلومات على هذا النحو بين الشركات التجارية.

كلما استخدمنا أحد محركات البحث، أُنشئت سجلاتٌ مهمتها تسجيل المواقع المُوصَى بها التي زرناها. وتحتوي هذه السجلات على معلومات مفيدة على غرار الكلمة المُستعلَم عنها نفسها، وعنوان آي بي للجهاز المُستخدَم، ووقت إرسال الاستعلام، والمدة التي قضيناها في كل موقع، وترتيب زيارتنا لهذه المواقع — كل ذلك من دون الكشف عن هُوياتنا. علاوةً على ذلك، تُسجِّل «سجلات تدفق النقر» المسارَ الذي سلكناه عند زيارتنا لمختلِف المواقع الإلكترونية، وكذلك تصفُّحنا لكل موقع. عندما نتصفَّح شبكة الويب، تُسجَّل كل نقرة ننقرها في مكانٍ ما لاستخدامها في المستقبل. البرامج المتوافرة للشركات تمكِّنها من جمع بيانات تدفق النقر التي تُنتجها مواقعها الإلكترونية — وتُعد هذه أداة تسويق لا تُقدَّر بثمن. على سبيل المثال، يمكن أن تساعد السجلات — من خلال ما تقدِّمه من بيانات عن النظام — في اكتشاف الأنشطة الضارة مثل سرقة الهُوية. كما يمكن استخدام

السجلات في قياس مدى فاعلية الدعاية عبر الإنترنت، وذلك بصفة أساسية عن طريق عدِّ مرات النقر على الإعلانات من قِبل زائري الموقع الإلكتروني.

من خلال تفعيل تحديد هُوية العميل، تُستخدَم ملفات تعريف الارتباط لإضفاء طابع شخصي على تجربة تصفُّحك. عندما تزور للمرة الأولى موقعًا إلكترونيًّا من اختيارك، سيُرسَل «ملف تعريف ارتباط»، وهو عبارة عن ملف نصي صغير يحتوي عادةً على مُعرِّف للموقع الإلكتروني ومُعرِّف للمستخدم، إلى جهاز الكمبيوتر لديك، إلا إذا حظرت استخدام ملفات تعريف الارتباط. وفي كل مرة تزور هذا الموقع الإلكتروني، يُرسِل ملفُّ تعريف الارتباط رسالةً إلى الموقع الإلكتروني، وبهذه الطريقة يظل يتعقَّب زياراتك. وكما سنرى في الفصل السادس، تُستخدَم ملفات تعريف الارتباط في تسجيل بيانات تدفُّق النقر، أو تعقُّب تفضيلاتك، أو إضافة اسمك إلى الإعلانات المستهدفة.

تُنتج مواقع شبكات التواصل الاجتماعي أيضًا كميات كبيرةً من البيانات، وفي هذا الصدد يأتي كلٌّ من فيسبوك وتويتر على رأس القائمة. بحلول منتصف عام ٢٠١٦، بلغ عدد مستخدمي فيسبوك، في المتوسط، ١٫٧١ مليار مستخدم نشط شهريًّا، جميعهم يُنتجون بيانات، ما نتج عنه حوالي ١٫٥بيتابايت (أو ١٠٠٠تيرابايت) من بيانات سجلات الويب يوميًّا. كان لموقع يوتيوب، موقع مشاركة مقاطع الفيديو الشهير، تأثيرٌ كبير منذ إطلاقه عام ٢٠٠٥، ويزعم بيان صحفي حديث عن يوتيوب أن عدد مستخدميه قد تجاوز المليار مستخدم في جميع أنحاء العالم. يمكن استخدام البيانات القيِّمة الناتجة عن محركات البحث ومواقع شبكات التواصل الاجتماعي في مجالاتٍ أخرى كثيرة، على سبيل المثال، عند التعامل مع المشكلات الصحية.

بيانات الرعاية الصحية

إذا تناولنا الرعاية الصحية، فسنجد أننا بصدد مجال يتضمَّن نسبةً كبيرة ومتزايدة من سكان العالم وهو آخذٌ في التحوُّل إلى نظام الحوسبة. تتحوَّل السجلات الصحية الإلكترونية تدريجيًّا لتصبح النظام المعتمد في المستشفيات وعيادات الأطباء، والهدف الأساسي من ذلك هو تسهيل مشاركة بيانات المرضى مع مستشفيات وأطباء آخرين؛ ومن ثمَّ تيسير توفير رعاية صحية أفضل. يتزايد جمع البيانات الشخصية عبر أجهزة الاستشعار القابلة للارتداء أو الزرع، لا سيَّما فيما يتعلق بالمتابعة الصحية، حيث أصبح الكثير منا يستخدمون أجهزةً لمتابعة اللياقة البدنية الشخصية متباينة التعقيد، والتي تُنتج المزيد من فئات البيانات. أصبح من الممكن الآن متابعة صحة المريض عن بُعد، وفي الوقت الحقيقي من خلال جمع البيانات عن ضغط الدم، ومعدل النبض، ودرجة حرارة الجسم، الأمر الذي ربما يقلِّل من تكاليف الرعاية الصحية ويحسِّن من جودة الحياة. تزداد أجهزة المتابعة عن بُعد هذه تطوُّرًا يومًا بعد يوم، وأصبحت الآن تتخطَّى القياسات الأساسية لتشمل متابعة النوم ومُعدَّل تشبُّع الشرايين بالأكسجين.

تُقدِّم بعضُ الشركات عوامل تحفيز لإقناع الموظفين باستخدام أجهزة اللياقة البدنية القابلة للارتداء، وتحقيق أهداف معينة مثل خسارة الوزن أو السير لعدد محدَّد من الخطوات كل يوم. وفي مقابل

الحصول على الجهاز، يوافق الموظف على مشاركة البيانات مع صاحب العمل. قد يبدو هذا الأمر منطقيًّا، ولكن ستنشأ حتمًا مشكلاتٌ تتعلَّق بالخصوصية لا بد من وضعها في الاعتبار، بالإضافة إلى الضغط غير المستحب الذي قد يشعر به البعض جراء الاشتراك في هذا النظام.

أصبحنا نرى بصورة متزايدة أشكالًا أخرى من متابعة الموظفين، مثل تتبُّع جميع أنشطة الموظفين على أجهزة الكمبيوتر والهواتف الذكية التي توفِّرها الشركة. وباستخدام برامج مخصَّصة، يمكن أن تشمل هذه المتابعة كلَّ شيء، بدءًا من متابعة المواقع الإلكترونية التي يجري تصفُّحها، ووصولًا إلى تسجيل عدد مرات الضغط على المفاتيح لكل موظف، والتحقق ممَّا إذا كان الحاسوب يُستخدَم لأغراض شخصية مثل تصفُّح مواقع شبكات التواصل الاجتماعي. في عصر التسريبات الهائلة للبيانات، أصبح الأمان هاجسًا متزايد الأهمية، ومن ثمَّ أصبح من الضروري حماية البيانات المؤسسية. وفي النهاية، فإن مراقبة رسائل البريد الإلكتروني وتتبُّع المواقع الإلكترونية التي جرت زيارتها مجرد طريقتَين للحد من سرقة المواد الحساسة.

رأينا بالفعل أنه يمكن استخراج البيانات الصحية الشخصية من أجهزة الاستشعار، مثل أجهزة متابعة اللياقة البدنية أو أجهزة متابعة الحالة الصحية. ولكن، الكثير من البيانات التي تُجمع من أجهزة الاستشعار هذه تُخصَّص لأغراض طبية عالية التخصص. إن بعضًا من أكبر مخازن البيانات الموجودة يجري إنشاؤه بالتزامن مع دراسة الباحثين لجينات العديد من الأنواع وتسلسل الجينوم لديها. شُرِحَت بنية جزيء الحمض النووي (دي إن إيه)، الذي يشتهر باحتوائه على التعليمات الوراثيةَ اللازمة لحياة الكائنات الحية، للمرة الأولى بوصفه حلزونًا مزدوجًا من قِبل جيمس واتسون وفرانسيس كريك عام ١٩٥٣. كان مشروع الجينوم البشري الدولي أحد أكثر المشروعات البحثية انتشارًا في السنوات الأخيرة، والذي يحدِّد التسلسل، أو الترتيب الدقيق، لثلاثة مليارات زوج من القواعد التي يتكوَّن منها الحمض النووي البشري. وفي نهاية المطاف، تساعد هذه البيانات الفِرق البحثية في دراسة الأمراض الوراثية.

البيانات في الوقت الحقيقي

تُجمَع بعض البيانات، وتُعالج، وتُستخدم في الوقت الحقيقي. سمحت زيادة قوة المعالجة الحاسوبية بزيادة القدرة على معالجة هذه البيانات وإنتاجها بسرعة. يحمل زمن الاستجابة في هذه الأنظمة أهميةً كبيرة؛ ومن ثمَّ يجب معالجة البيانات بصورة آنية. على سبيل المثال، يستخدم نظام تحديد المواقع العالمي (جي بي إس) نظامًا من الأقمار الصناعية لمسح الأرض وإرسال كميات هائلة من البيانات في الوقت الحقيقي. ومن ثمَّ، تُعالِج أجهزة استقبال نظام تحديد المواقع العالمي، والتي قد تكون في سيارتك أو هاتفك الذكي («ذكي» هنا تشير إلى أن جهازًا ما، هاتفًا في هذه الحالة، له القدرة على الوصول إلى شبكة الإنترنت وتقديم عدد من الخدمات أو التطبيقات التي يمكن ربطها معًا)، إشارات الأقمار الصناعية هذه وتحسب موقعك، وتوقيتك، وسرعتك.

أصبحت هذه التكنولوجيا مُستخدمة الآن في تطوير السيارات التي من دون سائق أو الذاتية القيادة. وهذه التكنولوجيا مُستخدمة بالفعل في مناطق محددة ومتخصصة مثل المصانع والمزارع، وتطوَّرت على يد عدد من كبار المُصنِّعين، بما في ذلك شركات فولفو، وتِسلا، ونيسان. أجهزة الاستشعار وبرامج الكمبيوتر المشاركة في هذه التكنولوجيا تعمل على معالجة البيانات في الوقت الحقيقي، حتى توجِّه السيارة بصورة يُعتمد عليها إلى وجهتك، وتتحكَّم في حركتها بالنسبة إلى مستخدمي الطريق الآخرين. يتطلَّب هذا رسمًا مسبقًا لخرائط ثلاثية الأبعاد للطرق لاستخدامها؛ لأن أجهزة الاستشعار لا يمكنها التعامل مع الطرق غير الظاهرة على الخرائط. تُستخدم أجهزة الاستشعار الرادارية لمتابعة حركة المرور للسيارات الأخرى، وترسل البيانات إلى كمبيوتر تنفيذي مركزي خارجي يتحكَّم في السيارة. ويجب برمجة أجهزة الاستشعار على رصد الأشكال والتمييز، على سبيل المثال، بين طفل يعدو عابرًا الطريق وجريدة تطير عبره؛ أو رصد، مثلًا، مخطَّط لخط السير في حالات الطوارئ بعد وقوع حادث. ولكن، هذه السيارات لا يمكنها بعد الاستجابة بالشكل المناسب لجميع المشكلات التي تفرضها البيئة الدائمة التغيُّر المحيطة بها.

وقع حادث التصادم المميت الأول الذي تضمَّن سيارةً ذاتية القيادة عام ٢٠١٦، عندما لم يُبدِ السائقُ البشري أو الآلي استجابةً تجاه اعتراضِ سيارة أخرى طريقَ هذه السيارة، بمعنى أن أيًّا منهما لم يضغط على مكابح السيارة استجابةً لذلك. أشارت شركة تِسلا، المُصنِّعة للسيارة الذاتية القيادة، في خبر صحفي في يونيو عام ٢٠١٦ إلى «الملابسات الشديدة الندرة التي أحاطت بحادث التصادم». يُنبِّه نظام القيادة الآلية السائقين إلى أنْ يُبقوا أيديهم على مقود السيارة طوال الوقت، بل إنه يتحقَّق حتى من أنهم يفعلون ذلك. صرَّحت شركة تِسلا بأن هذا الحادث هو حادث التصادم المميت الأول الذي يقع بسبب نظام القيادة الآلية لديها خلال ١٣٠ مليون ميل من القيادة، مقارنةً بحادث مميت واحد كل ٩٤ مليون ميل تتسبَّب فيه القيادة العادية غير الآلية في الولايات المتحدة.

تشير التقديرات إلى أن كل سيارة ذاتية القيادة ستنتج في المتوسط ٣٠تيرابايت من البيانات يوميًّا، ويجب معالجة الكثير منها في الوقت الحقيقي تقريبًا. يأمل مجالٌ بحثي جديد، يُدعَى «تحليلات تدفق البيانات»، وهو يتخطَّى الطرق التقليدية للإحصاء ومعالجة البيانات، في تقديمه وسيلةً لحل هذه المشكلة المتعلِّقة بالبيانات الضخمة تحديدًا.

البيانات الفلكية

في شهر أبريل ٢٠١٤، قدَّر تقرير أعدَّته مؤسسة البيانات الدولية أنه بحلول عام ٢٠٢٠، سيصل حجم الكون الرقمي إلى ٤٤ تريليون جيجابايت (الجيجابايت الواحد يساوي ١٠٠٠ميجابايت)؛ أي حوالي ١٠ أضعاف حجمه عام ٢٠١٣. ثمة كمٌّ متزايدٌ من البيانات تُنتجه التلسكوبات. على سبيل المثال، التلسكوب الكبير جدًّا في تشيلي، وهو عبارة عن تلسكوبٍ ضوئي يتكوَّن فعليًّا من أربعة تلسكوبات، ينتج كلٌّ منها كمًّا هائلًا من البيانات — ١٥تيرابايت كلَّ ليلة، وهذا كمُّ البيانات الإجمالي في الليلة الواحدة. يُعَد هذا التلسكوب حجر الأساس لمشروع المسح الشامل الكبير، وهو مشروع

يمتد لعشر سنوات يُنتِج بصورة متكرِّرة خرائط لسماء الليل، ويُقدَّر أنه سيُنتِج إجمالي ٦٠بيتابايت (الأصوات المُعطاة لكل صفحة ٢٠^٥بايت) من البيانات.

يوجد تلسكوبٌ أكبر من حيث إنتاج البيانات، وهو التلسكوب الراديوي «مصفوفة الكيلومتر المربع باثفيندر»، الذي أنشئ في أستراليا وجنوب أفريقيا، وبدأ العمل به عام ٢٠١٨. أنتجَ هذا التلسكوب ١٦٠تيرابايت من البيانات الخام كلَّ ثانية في بداية عمله، وازداد هذا الكمُّ مع اكتمال مراحله التالية. لن تُخزَّن جميع هذه البيانات، ولكن ستطرأ الحاجة إلى أجهزة كمبيوتر خارقة في جميع أنحاء العالم لتحليل البيانات المتبقية.

فيمَ تُستخدَم كلُّ هذه البيانات؟

من المستحيل تقريبًا في العصر الحالي أن يشارك المرء في الأنشطة اليومية ويتجنَّب ما يتم من جمع لبياناته الشخصية إلكترونيًّا. طاولات الدفع في المتاجر تجمع بياناتٍ عمَّا نشتريه، وشركات الطيران تجمع معلوماتٍ عن ترتيبات أسفارنا عندما نشتري تذكرة؛ والبنوك تجمع بياناتنا المالية.

تُستخدم البيانات الضخمة على نحو مُكثَّف في التجارة والطب، ولها تطبيقاتٌ في القانون، وعلم الاجتماع، والتسويق، والصحة العامة، وجميع فروع العلوم الطبيعية. للبيانات، بجميع صورها، القدرة على تقديم ثروة من المعلومات المفيدة إذا ما تمكنّا من ابتكار طرق لاستخراج تلك المعلومات. إن الأساليب الجديدة التي تمزج بين طرق الإحصاء التقليدية وعلوم الكمبيوتر تزيد من إمكانية التطبيق العملي لتحليل مجموعات البيانات الضخمة. طُوِّرَت هذه الأساليب والخوارزميات على أيدي إحصائيين وعلماء كمبيوتر يبحثون عن أنماط متكرِّرة في البيانات. ويُعَد تحديد الأنماط المهمة مفتاحَ نجاح عمليات تحليل البيانات الضخمة. كما أن التغيرات التي جلبها العصر الرقمي غيَّرت إلى حدٍّ كبير طُرق جمع البيانات، وتخزينها، وتحليلها. ومنحتنا ثورة البيانات الضخمة السيارات الذكية وأجهزة المراقبة المنزلية.

نتجَ عن القدرة على جمع البيانات إلكترونيًّا ظهور مجال علم البيانات المثير، الذي يجمع بين مجالَي الإحصاء وعلوم الكمبيوتر؛ من أجل تحليل هذه الكميات الكبيرة من البيانات لاكتشاف معارف جديدة في مجالات التطبيق المتعدِّدة الاختصاصات. إنَّ الهدف المطلق للعمل على البيانات الضخمة هو استخراج المعلومات المفيدة. وأصبح اتخاذ القرارات في الشركات يعتمد على نحو متزايد على المعلومات المستخرَجة من البيانات الضخمة، ومن المتوقَّع أن يزداد الاعتماد عليها أكثر في المستقبل. ولكن، ثمة مشكلات كبيرة، لا سيَّما في ظل قلة عدد علماء البيانات المُدرَّبين القادرين على تطوير الأنظمة اللازمة لاستخراج المعلومات المرغوبة وإدارتها على نحو فعَّال.

من خلال الاستعانة بطرق جديدة مستقاة من علم الإحصاء، وعلوم الكمبيوتر، والذكاء الاصطناعي، يجري الآن تصميم خوارزمياتٍ تقدِّم أفكارًا وتطويراتٍ جديدةً في مجال العلوم. على سبيل المثال، على الرغم من أنه لا يمكن توقُّع وقت حدوث الزلازل ومكانه، فإن عددًا متزايدًا من المؤسسات

تستخدم البيانات المُجمَّعة بواسطة الأقمار الصناعية وأجهزة الاستشعار الأرضية لمراقبة النشاط الزلزالي. والهدف من ذلك هو تحديد المكان التقريبي الذي من «المرجَّح» أن يشهد حدوث زلازل كبيرة على المدى الطويل. على سبيل المثال، قدَّرت هيئة المسح الجيولوجي الأمريكية، إحدى كبار المساهمين في أبحاث الزلازل، عام ٢٠١٦، أن «ثمة احتمالية قدرها ٧٦ في المائة أن زلزالًا شدَّتُه سبع درجات سيحدث في غضون الثلاثين عامًا القادمة في شمال كاليفورنيا». تساعد مثل هذه الاحتمالات في تكريس الموارد لوضع إجراءات، على غرار تحسين قدرة المباني على تحمُّل الزلازل، ووضع برامج لإدارة الكوارث وإدخالها حيِّز التنفيذ. تعمل العديد من الشركات، العاملة في هذه المجالات ومجالات أخرى، على البيانات الضخمة لتقديم أساليب تنبُّؤ مُحسَّنة، لم تكن متوافرةً قبل ظهور البيانات الضخمة. ومن ثمَّ، أصبحنا بحاجة إلى إلقاء نظرة على ما يميِّز البيانات الضخمة.

الفصل الثاني

لماذا البيانات الضخمة مميَّزة؟

لم تنشأ البيانات الضخمة من العَدَم؛ فهي وثيقة الصلة بتطور تكنولوجيا الكمبيوتر. أدَّى معدل النمو السريع للقدرات الحاسوبية وسعات التخزين إلى جمع كميات أكبر من البيانات مع الوقت، وبغض النظر عمَّن كان أول مَن صاغ مصطلح «البيانات الضخمة»، فإن الأمر كان يتعلَّق في البداية بالحجم فقط. ولكن، لا يمكن أن نقصُر تعريف البيانات الضخمة على عدد البيتابايت، أو حتى الإكسابايت، التي تُنتَج وتُخزَّن. ومع ذلك، فإن إحدى الوسائل المفيدة للحديث عن «البيانات الضخمة»، الناتجة عن انفجار البيانات، يقدِّمها مصطلح «البيانات الصغيرة»، وإن كان هذا المصطلح غير شائع الاستخدام بين جموع الإحصائيين. ولا شك أن مجموعات البيانات الضخمة كبيرة ومعقَّدة، ولكن، لكي نتوصَّل إلى تعريف لها، علينا أولًا أن نتعرَّف على «البيانات الصغيرة» ودورها في التحليل الإحصائي.

البيانات الضخمة في مقابل البيانات الصغيرة

عام ١٩١٩، وصل رونالد فيشر، الذي أصبح يشتهر الآن بكونه مؤسِّس علم الإحصاء الحديث بوصفه منهجًا أكاديميًّا دقيقًا، إلى محطة روتهامستد التجريبية الزراعية في المملكة المتحدة ليتولَّى مهمة تحليل بيانات المحاصيل الزراعية. جُمعَت البيانات من التجارب الميدانية الكلاسيكية التي أُجريت في روتهامستد منذ أربعينيات القرن التاسع عشر، بما في ذلك أبحاثها على القمح الشتوي والشعير الربيعي، وبيانات الأرصاد الجوية من المحطة الميدانية. أطلق فيشر مشروع «برودبالك» الذي درسَ تأثيرات الأسمدة المختلفة على القمح، ولا يزال هذا المشروع جاريًا حتى الآن.

حالما أدرك فيشر الحالة الفوضوية التي كانت عليها البيانات، اشتهر عنه أنه وصفَ بحثه الأولي هناك بأنه «التخلص من كومة الوحل». ولكن، من خلال الدراسة المدققة لنتائج التجارب التي كانت مسجَّلةً بعناية في دفاتر ملاحظات ذات أغلفة جلدية، تمكَّن فيشر من فهم ما تعنيه البيانات. كان فيشر يعمل وفقًا لإمكانات عصره المحدودة، قبل ظهور التكنولوجيا الحاسوبية المعاصرة، ولم يساعده إلا آلة حاسبة ميكانيكية في إجراء الحسابات، بشكل صحيح رغم ذلك، على البيانات المتراكمة على مدار ٧٠ عامًا. كانت هذه الآلة الحاسبة، المعروفة باسم «المليونير»، والتي كانت تعتمد في عملها على عملية تدوير شاقة لذراع يدوي، هي أحدث ابتكارات ذلك العصر؛ فقد كانت الآلةَ الحاسبة الأولى المتاحة تجاريًّا التي يمكن استخدامها لإجراء عملية الضرب. كان عمل فيشر مليئًا بالحسابات، ولعبت الآلة الحاسبة «المليونير» دورًا مهمًّا في تمكينه من إجراء العمليات الحسابية الكثيرة التي يمكن لأي كمبيوتر حديث إجراؤها في غضون ثوانٍ.

على الرغم من أن فيشر رتّب الكثير من البيانات وحللها، فإنها لا تَعد كميةً كبيرة بالمفهوم المعاصر، وممَّا لا شك فيه أنها لا تُعَد «بيانات ضخمة». كان جوهر عمل فيشر هو استخدام تجارب محدَّدة بدقة ومُراقبَة بعناية، ومُصمَّمة لإنتاج عينات بيانات عالية التنظيم وغير متحيِّزة. كان أسلوب العمل هذا ضروريًّا؛ لأنه لم يكن من الممكن تطبيق الأساليب الإحصائية التي توافرت في ذلك الوقت إلا على البيانات الهيكلية. ولا شك أن هذه الأساليب القيِّمة لا تزال تُمثِّل حجر الأساس لتحليل مجموعات البيانات الهيكلية الصغيرة. ولكن، لا يمكن تطبيق هذه الأساليب على كميات البيانات الكبيرة جدًّا التي أصبحنا قادرين على الوصول إليها حاليًّا عبر الكثير من المصادر الرقمية المختلفة المتاحة لنا.

تعريف البيانات الضخمة

في العصر الرقمي، لم نعد نعتمد بالكامل على العينات؛ وذلك لأننا أصبحنا قادرين على جمع كل البيانات التي نحتاجها عن شعوب بأكملها. إلا أن حجم هذه المجموعات من البيانات التي تزداد ضخامةً لا يمكنه بمفرده أن يقدِّم تعريفًا لمصطلح «البيانات الضخمة»، فعلينا أن نُدرج «التعقيد» في أيٍّ من تعريفاتها. وبدلًا من العينات المُعدَّة بعناية من «البيانات الصغيرة»، أصبحنا نتعامل الآن مع كميات هائلة من البيانات التي لم تُجمَع للإجابة عن أي أسئلة مطروحة، والتي تكون غير هيكلية عادة. من أجل توصيف السمات الرئيسية التي تجعل البيانات ضخمة، والاقترابِ من وضع تعريف للمصطلح، اقترح دوج لاني، في مقال كتبه عام ٢٠٠١، استخدام خصائص البيانات الضخمة الثلاث: الحجم، والتنوع، والسرعة. وبتناول كل عنصر من هذه العناصر على حدة، يمكننا تكوين فكرة أفضل عمَّا يعنيه مصطلح «البيانات الضخمة».

الحجم

يشير «الحجم» إلى كم البيانات الإلكترونية التي تُجمَع وتُخزَّن في الوقت الحالي، والذي ينمو بمعدَّل متزايد. البيانات الضخمة ضخمة بالفعل، ولكن ما مدى ضخامتها؟ قد يكون من السهل تحديد حجم معين لِمَا تعنيه كلمة «ضخمة» في هذا السياق، ولكن، ما كان يُعَد «ضخمًا» في الماضي، لم يَعُد ضخمًا بمعايير العصر الحالي. أصبح الحصول على البيانات يتزايد بمعدل ترتفع معه وتيرة التقادم الحتمي لأي حد نختاره. عام ٢٠١٢، أعلنت شركة آي بي إم وجامعة أكسفورد عن نتائج استطلاع رأي عن عمل البيانات الضخمة. في هذا الاستطلاع الدولي الذي شارك فيه ١١٤٤ مختصًّا يعملون في ٩٥ دولةً مختلفة، قال أكثر من نصفهم إن مجموعات البيانات التي يتراوح حجمها ما بين ١تيرابايت و١بيتابايت تُعد ضخمة، بينما جاء حوالي ثُلث المشاركين في فئة «لا أعلم». طلب الاستطلاعُ من المشاركين أن يختاروا سمةً أو اثنتَين من السمات المميِّزة للبيانات الضخمة من بين ثماني سمات، وصوَّتت نسبة ١٠ بالمائة فقط من المشاركين لسمة «الأحجام الكبيرة للبيانات»، في

حين كانت السمة الأكثر اختيارًا هي «نطاق أكبر من المعلومات»، والتي اجتذبت نسبة ١٨ بالمائة من المشاركين. السبب الآخر لعدم إمكانية وجود حد معين بناءً على الحجم فقط، هو أن ثمة عوامل أخرى، مثل سعة التخزين ونوع البيانات التي تُجمع، تتغيَّر بمرور الزمن، وتؤثِّر على إدراكنا للحجم. ولا شك أن بعض مجموعات البيانات ضخمة للغاية بالفعل، ومن بينها، على سبيل المثال لا الحصر، مجموعات البيانات الناتجة عن مصادم الهدرونات الكبير في مختبر سِرن، وهو مسارع الجسيمات الأول في العالم، والذي بدأ عمله عام ٢٠٠٨. حتى بعد استخراج نسبة واحد بالمائة فقط من إجمالي البيانات المُنتَجة، سيظل لدى العلماء ٢٥بيتابايت من البيانات ليعملوا على معالجتها سنويًّا. بوجهٍ عام، يمكننا القول إن معيار الحجم يمكن تلبيته إذا كانت مجموعة البيانات لا يمكن جمعها، وتخزينها، وتحليلها باستخدام أساليب الحوسبة والإحصاء التقليدية. تُعد بيانات الاستشعار، مثل تلك الناتجة عن مصادم الهدرونات الكبير، نوعًا واحدًا من البيانات الضخمة؛ ولذا دعونا نتناول بعضًا من الأنواع الأخرى.

التنوُّع

على الرغم من أنك قد ترى مصطلحَي «الإنترنت» و«شبكة الإنترنت العالمية» يُستخدمان عادةً على نحو متبادل، فإنهما في الحقيقة مختلفان تمامًا. الإنترنت عبارة عن شبكة من الشبكات، تتكوَّن من أجهزة كمبيوتر، وشبكات كمبيوتر، وشبكات مناطق محلية، وأقمار صناعية، وهواتف خلوية، وغيرها من الأجهزة الإلكترونية، جميعها متصلة معًا وقادرة على إرسال حِزَم من البيانات فيما بينها، ويُمكنها فعل ذلك باستخدام عنوان آي بي (بروتوكول الإنترنت). أمَّا شبكة الإنترنت العالمية (www أو الويب)، فيصفها مخترعها تي جيه بيرنرز لي بأنها «نظام معلومات عالمي» استغلَّ الاتصال بشبكة الإنترنت ليتمكَّن كلُّ مَن يملك جهازَ كمبيوتر واتصالًا بالإنترنت من التواصل مع مستخدمين آخرين عبر وسائط على غرار البريد الإلكتروني، والرسائل الفورية، وشبكات التواصل الاجتماعي، والرسائل النصية. ويمكن للمشتركين مع أحد مزوِّدي خدمات الإنترنت الاتصال بشبكة الإنترنت؛ ومن ثمَّ الوصول إلى الويب والكثير من الخدمات الأخرى.

بمجرد اتصالنا بالويب، يصبح لدينا وصول إلى مجموعة غير منظَّمة من البيانات، من مصادر موثوقة ومشبوهة، تكون عُرضةً للتكرار والخطأ. وهذا بعيد كل البُعد عن البيانات المرتَّبة الدقيقة التي تتطلَّبها أساليب الإحصاء التقليدية. على الرغم من أن البيانات المُجمَّعة من الويب يمكن أن تكون هيكلية، أو غير هيكلية، أو شبه هيكلية؛ ما ينتج عنه تنوُّع كبير (مثل مستندات معالجة النصوص أو منشورات مواقع شبكات التواصل الاجتماعي غير الهيكلية؛ وجداول البيانات شبه الهيكلية)، فإن أغلب البيانات الضخمة المستقاة من الويب تكون غير هيكلية. ينشر مستخدمو تويتر، على سبيل المثال، حوالي ٥٠٠ مليون رسالة مكوَّنة من ١٤٠ حرفًا كحدٍّ أقصى، أو «تغريدة»، كل يوم على مستوى العالم. تحمل هذه الرسائل القصيرة قيمةً تجارية عالية، وغالبًا ما تُحلَّل حسب إذا ما كانت المشاعر المُعبَّر عنها إيجابية، أم سلبية، أم محايدة. هذا المجال الجديد لتحليل المشاعر يتطلَّب أساليب مطوَّرةً بأسلوب خاص، وهو شيءٌ لا يمكن أن نؤديه بفاعلية إلا باستخدام تحليلات البيانات

الضخمة. على الرغم من التنوُّع الكبير للبيانات التي تجمعها المستشفيات، والجيش، والكثير من الشركات التجارية لأغراضٍ عدَّة، فإنه يمكن تصنيفها جميعها في نهاية المطاف بأنها هيكلية، أو غير هيكلية، أو شبه هيكلية.

السرعة

تتدفَّق البيانات في العصر الحالي باستمرار من مصادر على غرار الويب، والهواتف الذكية، وأجهزة الاستشعار. والسرعة ترتبط حتمًا بالحجم؛ كلما زادت سرعة إنتاج البيانات، زادت كميتها. على سبيل المثال، تنتقل الرسائل، التي أصبحت «تنتشر بسرعة»، على شبكات التواصل الاجتماعي بطريقة تجعل لها تأثير كُرة الثلج؛ أنشرُ شيئًا على إحدى شبكات التواصل الاجتماعي، ويراه أصدقائي، ويشاركه كلٌّ منهم مع أصدقائه، وهكذا. وتنتشر هذه الرسائل في جميع أنحاء العالم بسرعة كبيرة للغاية.

تُشير السرعة أيضًا إلى السرعة التي تُعالَج بها البيانات إلكترونيًّا. على سبيل المثال، من الضروري أن تُنتج بيانات الاستشعار، على غرار البيانات الناتجة عن السيارات الذاتية القيادة، في الوقت الحقيقي. فمن أجل أن تعمل السيارة بكفاءة، يجب أن تُحلَّل البيانات، التي تُرسَل لا سلكيًّا إلى موقع مركزي، بسرعة كبيرة للغاية حتى يمكن إرسال التعليمات الضرورية مرةً أخرى إلى السيارة على نحو آنٍ.

يمكن اعتبار التباين بُعدًا إضافيًّا لمفهوم السرعة؛ فهو يشير إلى معدلات التغيُّر في تدفق البيانات، مثل الزيادة الكبيرة في تدفق البيانات خلال أوقات الذروة. ويُعد هذا البُعد مهمًّا لأن أنظمة الكمبيوتر أصبحت حاليًّا أكثر عُرضةً للتعطُّل.

الموثوقية

بالإضافة إلى العناصر الثلاثة التي اقترحها لاني، يمكننا إضافة «الموثوقية» بوصفها العنصر الرابع. وتشير الموثوقية إلى جودة البيانات الجاري جمعها. كانت البيانات الدقيقة والموثوقة هي السمة المميِّزة للتحليل الإحصائي خلال القرن الماضي. وكان فيشر وغيره يَتُوقون إلى ابتكار أساليب تتضمَّن هذَين المفهومَين، إلا أن البيانات التي تُنتج في العصر الرقمي عادةً ما تكون غير هيكلية، وعادةً ما تُجمَع دون تصميم تجريبي، أو، بالتأكيد، دون أي فكرة عن الأسئلة التي قد تكون محور الاهتمام. ولكننا نسعى إلى الحصول على معلومات من هذا المزيج. لنتناوَل مثالًا على ذلك البيانات التي تُنتجها مواقع شبكات التواصل الاجتماعي. هذه البيانات، بطبيعتها، ليست دقيقة، أو موثوقة، وعادةً لا تكون المعلومات المنشورة صحيحة. كيف نثق إذن في أن البيانات تعطي نتائج ذات معنًى؟ يمكن أن يساعد الحجم في التغلُّب على هذه المشكلات، مثلما رأينا في الفصل الأول،

عندما وصف ثيوسيديدز استعانة القوات البلاتية بأكبر عدد ممكن من الجنود لِعَدِّ الطوب من أجل زيادة أرجحية تخمين الارتفاع الصحيح (التقريبي) للجدار الذي رغبوا في تسلُّقه. ولكن، علينا أن نكون أكثر حذرًا، حيث نعلم من النظرية الإحصائية أن زيادة الحجم يمكن أن تؤدِّي إلى نتائج عكسية؛ وذلك لأنه حتى في ظل وجود كمية كافية من بيانات، يمكننا العثور على عددٍ كبير من العلاقات الزائفة.

التمثيل المرئي والخصائص الأخرى

ظهر العديد من خصائص البيانات الضخمة الأخرى التي تنافست فيما بينها لتُضاف إلى خصائص البيانات الضخمة الثلاث الأصلية التي اقترحها لاني أو تحل محلها، مثل «قابلية التعرُّض للخطر» و«قابلية التطبيق»، ولعل أهم هذه الخصائص الإضافية «القيمة» و«التمثيل المرئي». تُشير القيمة بوجه عام إلى جودة النتائج المُستخرجة من تحليل البيانات الضخمة. كما أنها تُستخدَم لوصف عمليات بيع البيانات من قِبل الشركات التجارية إلى الشركات التي تتولَّى معالجتها باستخدام أساليب التحليل لديها؛ ومن ثمَّ فالقيمة مصطلحٌ شائع الاستخدام في مجال الأعمال القائمة على البيانات.

لا يُعَد التمثيل المرئي أحد الخصائص المُحدِّدة للبيانات الضخمة، ولكنه مهمٌّ فيما يخصُّ عرض النتائج التحليلية والتعريف بها. زاد تطوُّر الشكل المألوف للمخططات الدائرية الثابتة ورسوم التمثيل البياني بالأعمدة، التي ساعدتنا في فهم مجموعات البيانات الصغيرة، لمساعدتنا في تفسير البيانات الضخمة مرئيًّا، إلا أن إمكانية تطبيقها محدودة. على سبيل المثال، تُقدِّم المخططات البيانية للمعلومات عرضًا أكثر تعقيدًا، ولكنها مخططات ثابتة. وبما أن البيانات الضخمة يُضاف إليها المزيد باستمرار، فإن أفضل التمثيلات المرئية لها تكون تفاعليةً مع المستخدم، ويحدِّثها منشئُها بصفة منتظمة. على سبيل المثال، عندما نستخدم نظام تحديد المواقع العالمي (جي بي إس) لتخطيط مسار رحلة بالسيارة، فإننا نتعامل مع تمثيلاتٍ رسومية تفاعلية إلى حدٍّ كبير، بناءً على البيانات المُرسَلة من الأقمار الصناعية، لتتبُّع موقعنا.

تُمثِّل هذه الخصائص الأربع الرئيسية للبيانات الضخمة مجتمِعة؛ الحجم، والتنوُّع، والسرعة، والموثوقية، تحديًا كبيرًا فيما يتعلق بإدارة البيانات. ويمكن فهم المميزات التي نتوقَّع الحصول عليها من مواكبة هذه التحديات والأسئلة التي نأمل في الإجابة عنها باستخدام البيانات الضخمة، من خلال التنقيب في البيانات.

التنقيب في البيانات الضخمة

«البيانات هي النفط الجديد»، عبارة أصبحت متداولةً على نطاق واسع بين روَّاد الصناعة، والتجارة، والسياسة، وتُنسَب عادةً إلى كلايف همبي، مُبتكِر بطاقة ولاء عملاء تيسكو، عام ٢٠٠٦.

وهي عبارة جذابة تشير إلى أن البيانات، على غرار النفط، ذات قيمة كبيرة للغاية، ولكن يجب معالجتها أولًا قبل أن تُحقِّق هذه القيمة. استُخدِمَت هذه العبارة في الأساس كحيلة تسويقية استخدمها مزوِّدو خدمات تحليل البيانات على أمل أن يتمكَّنوا من بيع منتجاتِهم عن طريق إقناع الشركات بأن البيانات الضخمة هي المستقبل. وقد تكون كذلك بالفعل، ولكن، ظلَّت هذه الاستعارة قائمةً حتى يومنا هذا. فبمجرد أن تحصل على النفط، تكون لديك سلعة قابلة للتسويق. ولكن، لا ينطبق ذلك على البيانات الضخمة، فإنك لن تُنتِج أيَّ شيءٍ ذي قيمة إلا إذا امتلكَت البيانات المناسبة. تُمثِّل الملكية مشكلة، وتُمثِّل الخصوصية مشكلة؛ وعلى النقيض من النفط، لا يبدو أن البيانات مورد غير متجدِّد. ولكن، استمرارًا لهذه الاستعارة الصناعية، فإن التنقيب في البيانات الضخمة هو مهمة استخراج معلوماتٍ مفيدة وقيمة من مجموعات البيانات الهائلة الحجم.

باستخدام طريقتَي التنقيب في البيانات وتعلُّم الآلة، وكذلك الخوارزميات، لن يكون من الممكن اكتشاف الأنماط غير المعتادة أو الحالات غير المألوفة في البيانات فحسب، بل سيكون من الممكن أيضًا توقعها. وللحصول على هذا النوع من المعرفة من مجموعات البيانات الضخمة، قد يكون تعلُّم الآلة، بإشراف أو دون إشراف، أحد الأساليب المستخدمة. ويمكن اعتبار تعلُّم الآلة الخاضع للإشراف شبيهًا إلى حدٍّ ما بالتعلُّم القائم على الأمثلة لدى البشر. باستخدام بيانات التدريب، حيث تكون الأمثلة الصحيحة موسومةً أو مميَّزة، ينشئ برنامج كمبيوتر ما قاعدةً أو خوارزمية لتصنيف أمثلة جديدة. وتُفحَص هذه الخوارزمية باستخدام بيانات الاختبار. على النقيض من ذلك، تستخدم خوارزميات التعلُّم دون إشراف بياناتٍ مُدخلَة غير موسومة ومن دون تحديد هدف معين؛ فهي مُصمَّمة لاستكشاف البيانات واكتشاف الأنماط الخفية.

وكمثال على ذلك، دعونا نتناول كشف الاحتيال المرتبط ببطاقات الائتمان، ونرى كيف تُستخدم كل طريقة.

كشف الاحتيال في استخدام بطاقات الائتمان

تُبذل الكثير من الجهود لاكتشاف الاحتيال في استخدام بطاقات الائتمان والحيلولة دون حدوثه. إذا كنت تعِسَ الحظ ووصلتكَ مكالمة من مكتب كشف احتيال بطاقات الائتمان التابع له، فقد تتساءل كيف اتُّخذ القرار بأن آخِر عملية شراء تمَّت باستخدام بطاقتك الائتمانية من المحتمل أن تكون ضربًا من الاحتيال. بالنظر إلى العدد الهائل للمعاملات التي تتم باستخدام بطاقات الائتمان، لم يعد من المناسب أن يتولَّى البشر فحص هذه العمليات باستخدام طرق تحليل البيانات التقليدية؛ ومن ثَمَّ أصبحت أدوات تحليل البيانات الضخمة على نحو متزايد ضرورةً لا غنى عنها. إن عزوف المؤسسات المالية عن مشاركة تفاصيل أساليبها للكشف عن الاحتيال ببطاقات الائتمان أمرٌ مفهوم؛ حيث إن ذلك سيمنح المجرمين الإلكترونيين المعلومات التي يحتاجونها لابتكار طرق للتحايل عليها. ولكن، يمكن أن نحصل على فكرة جيدة عن هذا الموضوع دون الخوض في تفاصيله الكاملة.

ثمة العديد من السيناريوهات المحتملة، ولكننا سنتناول الخدمات المصرفية الشخصية وسنستعرض حالات سرقة بطاقات الائتمان واستخدامها مع معلومات أخرى مسروقة، مثل رقم التعريف الشخصي للبطاقة (بي آي إن). في هذه الحالة، قد تُظهِر البطاقة زيادةً مفاجئة في الإنفاق، وهي عملية احتيال من السهل اكتشافها بواسطة الجهة المُصدِرة للبطاقة. وفي أغلب الأحيان، يستخدم المحتال البطاقة المسروقة للمرة الأولى في إجراء «معاملة تجريبية» حيث يشتري شيئًا غير باهظ الثمن. وإن لم تُثِر هذه المعاملة أي إنذارات، يبدأ في الاستيلاء على مبالغ أكبر. قد تنطوي هذه المعاملات على احتيال وقد لا تنطوي؛ فربما اشترى صاحب البطاقة شيئًا خارج نمط مشترياته المعتاد، أو ربما أنفق فعلًا الكثير من المال خلال هذا الشهر. كيف نكتشف إذن المعاملات التي تنطوي على احتيال؟ دعونا نتناول أولًا أسلوبًا دون إشراف يُدعى «التجميع»، وكيف يمكن استخدامه في مثل هذا الموقف.

التجميع

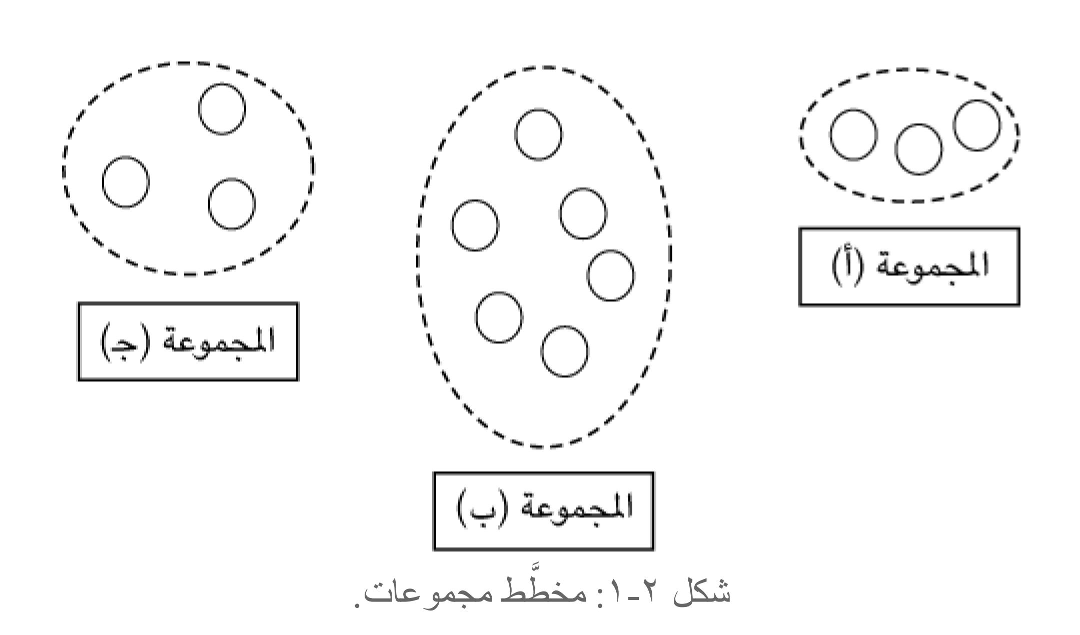
بِناءً على خوارزميات الذكاء الاصطناعي، يمكن استخدام أساليب التجميع في اكتشاف التضارب أو الانحراف في سلوكيات العملاء الشرائية. ويتحقَّق ذلك عن طريق البحث في بيانات المعاملات بغرض اكتشاف أي شيءٍ غير معتاد أو مشتبه فيه، والذي ربما يكون ضربًا من الاحتيال أو لا يكون.

تجمع شركات بطاقات الائتمان كمًّا كبيرًا من البيانات وتستخدمه في إنشاء ملفات بياناتٍ تعرض سلوكيات الشراء لدى عملائها. ومن ثمَّ، تُحدَّد مجموعاتٌ من ملفات البيانات ذات الخصائص المتماثلة إلكترونيًّا بواسطة برنامج كمبيوتر «تكراري» (أي يُكرِّر عمليةً ما حتى يصل إلى نتيجة معينة). على سبيل المثال، قد تُحدَّد مجموعة للحسابات طبقًا للموقع أو لنطاق الإنفاق المعتاد، أو طبقًا للحد الأعلى لإنفاق العميل، أو طبقًا لنوع السلع المُشتراة، وكلٌّ منها تتولَّد عنه مجموعة منفصلة.

عندما تُجمَع البيانات بواسطة أحد مزوِّدي خدمة بطاقات الائتمان فإنها لا تحمل أي وصف يشير إلى أن المعاملات مشروعة أو احتيالية. ومهمتنا هي استخدام هذه البيانات بوصفها مُدخلات، واستخدام إحدى الخوارزميات المناسبة، وتصنيف المعاملات بدقة. ولتحقيق ذلك، سنحتاج إلى البحث عن مجموعات، أو فئات، مماثلة ضمن بيانات المُدخلات. إذن، يمكننا أن نجمع البيانات، على سبيل المثال، طبقًا للمبالغ المُنفقَة، أو مكان إجراء المعاملة، أو نوع عملية الشراء، أو عُمر صاحب البطاقة. وعند إجراء معاملة جديدة، يُسجَّل رقم تعريف المجموعة لهذه المعاملة، وإذا كان مختلفًا عن رقم تعريف المجموعة الحالية للعميل، تُعتبر المعاملة مشتبهًا فيها. حتى وإن حلَّت المعاملة ضمن المجموعة المعتادة، فإنها تظل مثار شك إذا كانت بعيدةً بقدرٍ كافٍ عن مركز المجموعة.

على سبيلِ المثال، لنفترِض أن جدةً تبلغ من العُمر ٨٣ عامًا تعيش في باسادينا اشترت سيارةً رياضية جذَّابة، إذا لم تحلَّ عملية الشراء هذه ضمن مجموعة سلوكيات شرائها المعتادة، مثل البقالة

وزيارات مصفَف الشعر، فإنها تَعد انحرافًا. وأيُّ شيءٍ خارج عن المألوف، مثل عملية الشراء هذه، يُنظر إليه على أنه يستوجب مزيدًا من البحث والتحقيق، وهو ما يبدأ عادةً بالتواصل مع مالك البطاقة. في شكل ٢-١، نرى مثالًا بسيطًا للغاية على مخطط مجموعاتٍ يمثِّل هذه الحالة.



شكل ٢-١: مخطَّط مجموعات.

تُوضِّح المجموعة (ب) نفقات الجَدَّة الشهرية المعتادة مُجمَّعةً في مجموعة واحدة مع أشخاص آخرين ينفقون نفس المبالغ شهريًّا. ولكن، في بعض الحالات، كالحال عند حصولها على عطلتها السنوية، تزداد نفقات الجَدَّة خلال هذا الشهر، وربما تُوضَع في هذه الحالة مع الأشخاص في المجموعة (ج)، والتي لا تبعد كثيرًا عن المجموعة (ب)؛ ومن ثَمَّ، لا تُعد مختلفةً عنها إلى حدٍّ كبير. حتى وإن كان الأمر كذلك، بما أن هذه المصروفات حلَّت في مجموعة مختلفة، فسيتم التحقُّق منها بوصفها نشاطًا مشبوهًا للحساب، إلا أن شراءها للسيارة الرياضية الجذابة عبر حسابها يضع مصروفاتها في المجموعة (أ)، والتي تبعد كثيرًا عن مجموعتها المعتادة، وعليه، فمن غير المرجَّح أن تعكس عملية شراء مشروعة.

وعلى النقيض من هذه الحالة، إذا كان لدينا بالفعل مجموعة من الأمثلة التي نعلم يقينًا أن احتيالًا حدث خلالها، فبدلًا من خوارزميات التجميع، يمكننا استخدام أساليب التصنيف، التي من شأنها أن تمدَّنا بأسلوب آخر من أساليب التنقيب في البيانات، يُستخدم في الكشف عن الاحتيال.

التصنيف

التصنيف هو أحد أساليب التعلُّم الخاضع لإشراف، ويتطلَّب معرفةً مسبقة بمجموعات البيانات المستخدمة. في هذا الأسلوب، نبدأ بمجموعة بيانات تكون فيها كلُّ ملاحظة مضافًا إليها وصف أو

مُصنّفة على نحو صحيح بالفعل. وتنقسم مجموعة البيانات هذه إلى «مجموعة تدريب»، تَمكّننا من إنشاء نموذج تصنيف من البيانات، و«مجموعة اختبار»، تُستخدَم للتحقُّق من أن النموذج جيد. ثم يمكننا استخدام هذا النموذج في تصنيف ملاحظاتٍ جديدة حال ظهورها.

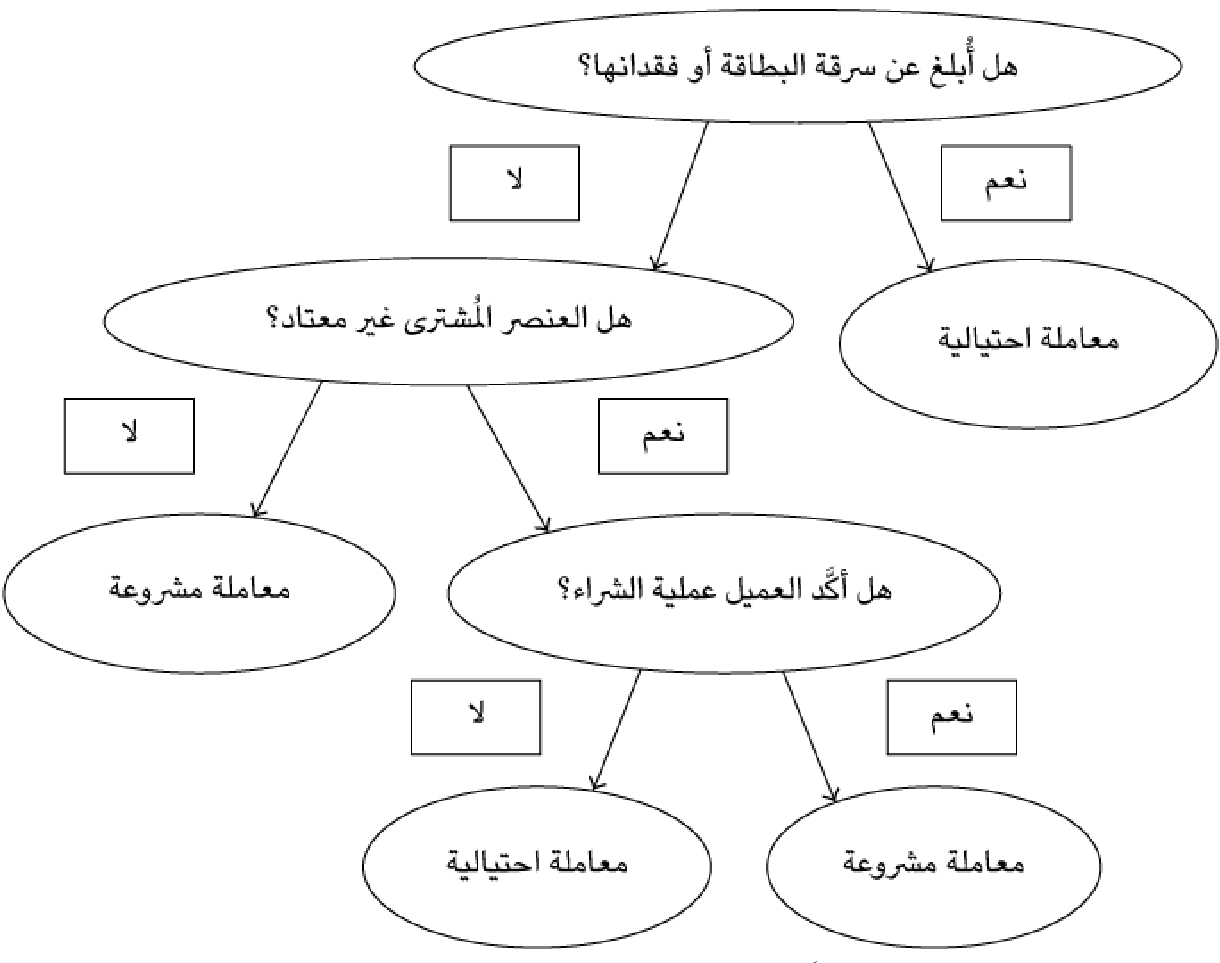
لتوضيح مفهوم التصنيف، سننشئ مخططًا صغيرًا لتسلسل اتخاذ القرارات؛ لكي نكشف الاحتيال باستخدام بطاقات الائتمان.

لإنشاء مخطط اتخاذ القرارات الشجري، دعونا نفترض أن بيانات معاملات بطاقة الائتمان قد جُمِعَت، وأن المعاملات صُنِّفت إلى مشروعة أو احتيالية بِناءً على معرفتنا السابقة، كما يظهر في جدول ٢-١.

باستخدام هذه البيانات، يمكننا إنشاء مخطَّط اتخاذ قراراتٍ شجري، كالموضَّح في شكل ٢-٢، والذي يمكِّن الكمبيوتر من تصنيف المعاملات الجديدة المُدخَلة إلى النظام. ونأمل أن نصل إلى أحد تصنيفَي المعاملات المحتمَلَين؛ إمَّا مشروعة أو احتيالية، عن طريق طرح مجموعة من الأسئلة.

جدول ٢-١: مجموعة بيانات احتيالية ذات تصنيفاتٍ معلومة

هل أُبلغ عن سرقة البطاقة أو فقدانها؟	هل العنصر المُشترى غير معتاد؟	هل تمَّ الاتصال بالعميل وسؤاله عمَّا إذا كان قد أجرى عملية الشراء هذه؟	التصنيف
لا	لا		معاملة مشروعة
لا	نعم	نعم	معاملة مشروعة
لا	نعم	لا	معاملة احتيالية
نعم			معاملة احتيالية



شكل ٢-٢: مخطَّط اتخاذ القرارات الشجري الخاص بالمعاملات.

بدءًا من قمة المخطط الشجري في شكل ٢-٢، نجد أن لدينا مجموعةً من الأسئلة الاختبارية التي ستمكِّننا من تصنيف المعاملات الجديدة.

على سبيل المثال، إذا أظهر حساب السيد سميث أنه أبلغَ عن فقدان بطاقة ائتمانه أو سرقتها، فإن أي محاولة لاستخدامها ستُعَدُّ احتيالًا. وإذا لم يُبلِغ عن فقدان البطاقة أو سرقتها، فإن النظام سيتحقَّق ممَّا إذا اشتُري عنصرٌ غير معتاد أو عنصرٌ يتكلَّف مبلغًا لم يَعتد هذا العميل إنفاقه. إذا لم يحدث ذلك، فلن تُعتبر المعاملة غير معتادة، وستُصنَّف بأنها مشروعة. من ناحية أخرى، إذا كان العنصر غير معتاد، فسيتلقَّى السيد سميث مكالمةً هاتفية. إذا أكَّد على أنه أجرى معاملة الشراء، فستُعَدُّ مشروعة؛ وإن لم يؤكِّد ذلك، فستُعَدُّ احتيالًا.

بعدما توصَّلنا إلى تعريفٍ غير رسمي للبيانات الضخمة، وسلَّطنا الضوءَ على أنماط الأسئلة التي يُمكن الإجابة عنها من خلال التنقيب في البيانات الضخمة، دعونا نتناول الآن المشكلات المتعلقة بتخزين البيانات.

الفصل الثالث

تخزين البيانات الضخمة

كانت سعة تخزين القرص الصلب الأول، الذي ابتكرته شركة آي بي إم وباعته في مدينة سان خوزيه بولاية كاليفورنيا، حوالي ميجابايت، وكان يحتوي على ٥٠ قرصًا يبلغ قطر كلٍّ منها ٢٤ بوصة. كان هذا القرص الصلب أحدث تقنية موجودة عام ١٩٥٦. كان حجم الجهاز هائلًا؛ فقد كان يزن ما يزيد عن الطن، وكان يمثِّل جزءًا من جهاز كمبيوتر مركزي. عند هبوط الرحلة أبوللو ١١ على سطح القمر عام ١٩٦٩، كان مركز وكالة ناسا لرحلات الفضاء المأهولة في هيوستن يستخدم أجهزة كمبيوتر مركزية، احتوى كلٌّ منها على ذاكرة مساحتها تصل إلى ٨ميجابايت. المثير للدهشة أن الكمبيوتر الذي كان على متن سفينة الفضاء أبوللو ١١ التي هبطت على سطح القمر، والتي كان يقودها نيل أرمسترونج، كان يحتوي على ذاكرة مساحتها ٦٤كيلوبايت فقط.

تطوَّرت تقنيات الكمبيوتر سريعًا، وبحلول بداية ازدهار أجهزة الكمبيوتر الشخصية في ثمانينيات القرن العشرين، كان متوسط حجم القرص الصلب في الكمبيوتر الشخصي ٥ميجابايت إذا كان الكمبيوتر يتضمَّن قرصًا صلبًا بالفعل؛ إذ لم يكن الحال على هذا المنوال دائمًا. وهذه المساحة تكفي لتخزين صورة أو صورتَين في يومنا هذا. زادت سعة تخزين أجهزة الكمبيوتر بسرعة كبيرة، وعلى الرغم من أن سعة تخزين الكمبيوتر الشخصي لم تواكب تخزين البيانات الضخمة، فإنها زادت على نحو كبير في السنوات الأخيرة. أصبح بإمكانك حاليًّا شراء كمبيوتر شخصي مساحة قرصه الصلب ثمانية تيرابايت أو أكثر. وأصبحت محركات الأقراص المحمولة متوافرةً حاليًّا بسعة تخزين ١تيرابايت، وهو ما يكفي لتخزين حوالي ٥٠٠ ساعة من الأفلام أو ما يزيد على ٣٠٠ ألف صورة. ستبدو هذه السعات كبيرةً حتى نقارنها بحجم البيانات الجديدة التي تُنتَج كلَّ يوم، والذي يُقدَّر بحوالي ٢٫٥إكسابايت.

عندما استُبدلَت الصمامات بالترانزستورات في ستينيات القرن العشرين، تنامَى عددُ الترانزستورات التي يمكن وضعها على شريحة إلكترونية واحدة بسرعة كبيرة للغاية، بما يتناسب مع قانون مور تقريبًا، والذي سنتناوله في الجزء الآتي من الكتاب. وعلى الرغم من التوقعات بأننا شارفنا على الوصول إلى الحد الأقصى للتصغير، فإن الأمر يظل مقاربةً معقولة ومفيدة. أصبح في مقدورنا الآن رصُّ مليارات الترانزستورات المتزايدة السرعة على شريحة واحدة، الأمر الذي يتيح لنا تخزين كميات أكبر من البيانات، في حين تسمح المعالِجات المتعدِّدة النوى، بالاشتراك مع برنامج كمبيوتر ذي مؤشرات ترابط متعدِّدة، بمعالجة هذه البيانات.

قانون مور

عام ١٩٦٥، تنبَّأ جوردون مور، الذي أصبح أحد الشركاء المؤسِّسين لشركة إنتل، بأنه على مدار السنوات العشر القادمة، سيتضاعف تقريبًا عدد الترانزستورات التي يمكن وضعها في شريحةٍ كلَّ ٢٤ شهرًا. وعام ١٩٧٥، غيَّر مور من تنبُّئه وقال إن التعقيد سيتضاعف كلَّ ١٢ شهرًا على مدار خمس سنوات، ثم عاد مرةً أخرى ليقول إنه سيتضاعف كلَّ ٢٤ شهرًا. اقترح ديفيد هاوس، وهو زميلٌ من شركة إنتل، بعد وضع السرعة المتزايدة للترانزستورات في الاعتبار، أن «أداء» الشرائح الإلكترونية المصغرة سيتضاعف كلَّ ١٨ شهرًا، وهذا التنبؤ الأخير هو الأكثر استخدامًا حاليًّا فيما يخصُّ قانون مور. أثبتَ هذا التنبؤ دقته البالغة؛ فقد أصبحت أجهزة الكمبيوتر حقًّا أسرع، وأرخص، وأقوى ممَّا كانت عليه عام ١٩٦٥، إلا أن مور نفسه يشعر بأن هذا «القانون» لن يستمر طويلًا.

طبقًا لِمَا كتبه إم ميتشل والدروب في مقاله الذي نُشِر في عدد شهر فبراير ٢٠١٦ من مجلة «نيتشر» العلمية، فقد اقتربت نهاية قانون مور. المعالج الدقيق هو الدائرة المتكاملة المسئولة عن تنفيذ التعليمات التي يقدِّمها برنامج الكمبيوتر. يتكوَّن هذا المعالج عادةً من مليارات الترانزستورات المُكدَّسة في مساحة صغيرة للغاية على شريحة مصغَّرة من السيليكون. وثمة بوابة في كل ترانزستور تسمح بتشغيله أو إلغاء تشغيله حتى يمكن استخدامه في تخزين صفر أو واحد. ويتدفَّق تيار دَخْل ضئيل للغاية عبر كل بوابة ترانزستور، ويُنتِج تيار خَرْج مُضَخَّم عند إغلاق البوابة. كان ميتشل والدروب مهتمًّا بالمسافة بين البوابات، وهي حاليًّا عبارة عن فجوات يبلغ حجم الواحدة منها ١٤ نانومترًا في أفضل أنواع المعالِجات الدقيقة، وصرَّح بأن مشكلات ارتفاع درجة الحرارة الناتجة عن تقارب الدوائر الإلكترونية، وكيف يمكن تشتيتها بفاعلية، تتسبَّب في تداعي النمو الأُسِّي الذي تنبَّأ به قانون مور، الأمر الذي لفت انتباهنا إلى الحدود القصوى الأساسية التي رأى أننا نقترب منها بسرعة.

النانومتر الواحد يساوي ١٠$^{-٩}$ متر، أو جزءًا من المليون من المليمتر. ولوضع هذا القياس ضمن سياق، يبلغ قطر الشعرة لدى الإنسان حوالي ٧٥ ألف نانومتر، ويتراوح قطر الذرة ما بين ٠,١ و٠,٥ نانومتر. زعم باولو جارجيني، أحد العاملين في شركة إنتل، أن الحد الأقصى للفجوات سيكون ٢ أو ٣ نانومترات، وأننا سنصل إلى هذا الحد الأقصى في المستقبل القريب، ربما حين ندخل عشرينيات القرن الحادي والعشرين. تنبَّأ والدروب أنه «بهذا المعدل، سيكون سلوك الإلكترونات محكومًا بمبدأ عدم اليقين الكَمِّي الذي سيجعل الترانزستورات غير موثوقة على نحو ميئوس منه». وكما سنرى في الفصل السابع، من المرجَّح جدًّا فيما يبدو أن أجهزة الكمبيوتر الكَمِّية، وهي تقنية لا تزال في مهدها، هي التي سترسم في نهاية المطاف الخطوات المقبلة في هذا الشأن.

لا يزال قانون مور قابلًا للتطبيق حتى يومنا هذا على معدَّل نمو البيانات؛ إذ يبدو أن كمية البيانات المُنتجة تتضاعف تقريبًا كلَّ عامَين. كما تزداد كمية البيانات بزيادة سعة التخزين وزيادة القدرة على معالجة البيانات. ونحن المستفيدون من ذلك: أصبحت نتفليكس، والهواتف الذكية، وإنترنت الأشياء (طريقة ملائمة لتسمية العدد الهائل من أجهزة الاستشعار الإلكترونية المتصلة بالإنترنت)، والحوسبة السحابية (شبكة عالمية من الخوادم المتصلة فيما بينها)، من بين خدماتٍ أخرى، ممكنةً بفضل النمو الأُسي الذي تنبَّأ به قانون مور. كلُّ هذه البيانات المُنتجة بحاجة إلى التخزين، وهذا ما سنتناوله فيما يلي.

تخزين البيانات الهيكلية

يمكن لأي شخص يستخدم كمبيوترًا شخصيًّا، أو كمبيوترًا محمولًا، أو هاتفًا ذكيًّا، الوصول إلى البيانات المُخزَّنة في قواعد البيانات. تُخزَّن البيانات الهيكلية، مثل كشوف الحسابات المصرفية وأدلة العناوين الإلكترونية، في قواعد بيانات ارتباطية. ومن أجل إدارة هذا الكَمِّ من البيانات الهيكلية، يُستخدَم نظام إدارة قواعد بيانات ارتباطية لإنشاء البيانات، والحفاظ عليها، والوصول إليها، ومعالجتها. تتمثَّل الخطوة الأولى في تصميم مخطط قاعدة البيانات (أي بنية قاعدة البيانات). ولتحقيق ذلك، علينا أن نعرف حقول البيانات، وأن نكون قادرين على تنظيمها في جداول، ومن ثَمَّ، سيكون علينا أن نحدِّد العلاقات بين الجداول. بمجرد الانتهاء من ذلك وإنشاء قاعدة البيانات، يمكننا أن نملأها بالبيانات وإجراء استعلاماتٍ فيها باستخدام لغة الاستعلام الهيكلية (SQL).

من الجليِّ أنه يجب تصميم الجداول بعناية، وقد يتطلَّب الأمرُ الكثيرَ من العمل لإجراء تغييراتٍ كبيرة. ولكن، ينبغي عدم التقليل من شأن النموذج الارتباطي. فمقارنةً بالكثير من تطبيقات البيانات الهيكلية، يُعد هذا النموذج سريعًا وموثوقًا. يتضمَّن أحد الجوانب المهمة لتصميم قواعد البيانات الارتباطية عمليةً تُسمَّى «التسوية»، وتشمل تقليل تكرار البيانات إلى الحد الأدنى؛ ومن ثَمَّ، الحد من متطلَّبات التخزين. وتسمح هذه العملية بإجراء استعلامات أسرع، ولكن، رغم ذلك، كلما زاد حجم البيانات تراجعَ أداء قواعد البيانات التقليدية هذه.

تكمن المشكلة في قابلية التوسُّع. بما أن قواعد البيانات الارتباطية صُمِّمت في الأساس لتعمل على خادم واحد فقط، فإن سرعتها وموثوقيتها تتراجعان كلما أُضيفَ المزيد من البيانات. الحل الوحيد لتحقيق قابلية التوسُّع هو إضافة المزيد من القدرة الحاسوبية، والتي لها حدٌّ أقصى أيضًا. يُعرَف هذا باسم «قابلية التوسُّع الرأسية». على الرغم من أن البيانات الهيكلية عادةً ما تُخزَّن وتُدار في نظام إدارة قواعد بياناتٍ ارتباطية، فإن كفاءة نظام إدارة قواعد البيانات الارتباطية تقل، حتى مع البيانات الهيكلية، عندما تكون البيانات ضخمة؛ أي عندما يكون حجمها بالتيرابايت أو البيتابايت أو أكثر.

من السمات المهمة لقواعد البيانات الارتباطية وأحد الأسباب الجيدة للاستمرار في استخدامها هو أنها تتماشى مع الخصائص الأربع الآتية: الذَّريَّة، والاتساق، والعزل، والاستمرارية. تضمن الذَّريَّة عدم تحديث قواعد البيانات بواسطة المعاملات غير الكاملة، ويستبعد الاتساق البيانات غير الصحيحة، ويضمن العَزل عدم تداخل معاملة مع أخرى، وتعني الاستمرارية ضرورة تحديث قاعدة البيانات قبل تنفيذ المعاملة التالية. جميع هذه الخصائص مُستحبَّة، إلا أن تخزين البيانات الضخمة، التي تكون في الغالب غير هيكلية، والوصول إليها، يتطلَّبان نهجًا مختلفًا.

تخزين البيانات غير الهيكلية

فيما يخصُّ البيانات غير الهيكلية، لا يصلح استخدام نظام إدارة قواعد البيانات الارتباطية لعدة أسباب، لا سيَّما أنه بمجرد إنشاء مخطط قاعدة البيانات الارتباطية، يُصبح من الصعب تغييره. علاوةً على ذلك، لا يمكن تنظيم البيانات غير الهيكلية في صفوف وأعمدة بما يحقِّق سهولة الاستخدام. وكما رأينا سابقًا، عادةً ما تكون البيانات الضخمة عالية السرعة وتُنتَج في الوقت الحقيقي وتتطلَّب معالجة آنية؛ ولذا على الرغم من أن نظام إدارة قواعد البيانات الارتباطية يصلح بامتياز للعديد من الأغراض ويفيدنا كثيرًا، فقد أجريت على ضوء انفجار البيانات الحالي أبحاثٌ مكثَّفة في أساليب التخزين والإدارة الجديدة.

لتخزين مجموعات البيانات الهائلة هذه، تُوزَّع البيانات على خوادم. وكلما زاد عدد الخوادم المُتضمَّنة، زادت أيضًا احتمالية حدوث عُطل في مرحلة ما، وعليه، فمن المهم أن تكون لدينا عدة نسخ متطابقة من البيانات نفسها، وتُخزَّن كلُّ نسخة على خادم مختلف. وممَّا لا شك فيه أنه في ضوء كميات البيانات الهائلة الجاري معالجتها حاليًّا، أصبح يُنظر إلى أعطال الأنظمة على أنها أمرٌ حتمي؛ ومن ثمَّ أصبحت طرق التغلُّب عليها مُضمَّنةً في أساليب التخزين. كيف تُلبَّى إذن متطلَّباتُ السرعة والموثوقية؟

نظام هادوب للملفات الموزَّعة

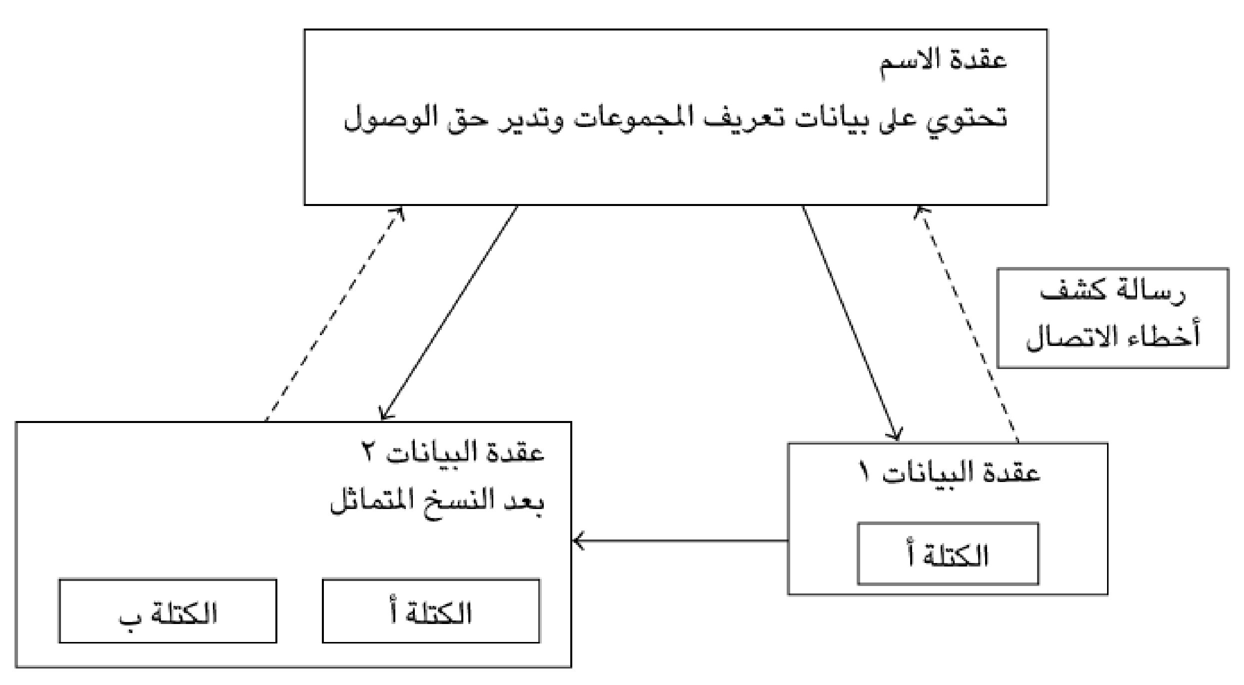
يوفِّر نظام الملفات الموزَّعة قدرةً تخزينية فعالة وموثوقة للبيانات الضخمة عبر الكثير من أجهزة الكمبيوتر. من منطلق تأثره بالأفكار التي نُشرَت في أكتوبر ٢٠٠٣ بواسطة شركة جوجل في ورقة بحثية عن إطلاق نظام ملفات جوجل، بدأ دوج كاتينج، الذي كان يعمل في ذلك الوقت في شركة ياهو، وزميله مايك كافاريلا، طالب الدراسات العليا في جامعة واشنطن، العمل على تطوير نظام هادوب للملفات الموزَّعة. يُعَد هادوب، وهو أحد أشهر أنظمة الملفات الموزَّعة، جزءًا من مشروع أكبر للبرامج المفتوحة المصدر يُسمَّى «هادوب إيكوسيستم». سُمي النظام باسم هادوب تيمُّنًا بدُمية صفراء محشوة على هيئة فيل كانت مملوكةً لابن كاتينج، والمشروع مكتوب بلغة البرمجة الشهيرة جافا. إذا كنت تستخدم فيسبوك، أو تويتر، أو إيباي، على سبيل المثال، فاعلم أن هادوب يعمل في الخلفية أثناء ذلك. يسمح النظام بتخزين البيانات شبه الهيكلية وغير الهيكلية، ويوفر منصةً لتحليل البيانات.

عندما نستخدم نظام هادوب للملفات الموزَّعة، تُوزَّع البيانات عبر الكثير من العُقَد التي يُقدَّر عددُها بنحو عشرات الآلاف، والموجودة فعليًّا في مراكز بيانات في جميع أنحاء العالم. يوضِّح شكل ٤ البنية الأساسية لمجموعة واحدة من نظام هادوب للملفات الموزَّعة، والتي تتكوَّن من عُقدة اسم رئيسية واحدة والكثير من عُقَد البيانات الفرعية.

تتعامل عُقدة الاسم NameNode مع جميع الطلبات التي تصل من كمبيوتر عميل، وتوزِّع مساحة التخزين، وتتابع المساحة المتوافرة للتخزين وموقع البيانات. كما أنها تدير جميع عمليات الملفات الأساسية (مثل فتح الملفات وإغلاقها) وتتحكَّم في الوصول إلى البيانات بواسطة أجهزة

الكمبيوتر العميل. أمَّا عُقَد البيانات DataNodes، فإنها تكون مسئولةً عن التخزين الفعلي للبيانات، ولفعل ذلك تنشئ الكتل وتحذفها وتنسخها حسب الضرورة.

يُعَد النسخ المتماثل للبيانات إحدى السمات الرئيسية لنظام هادوب للملفات الموزَّعة. على سبيل المثال، بالنظر إلى شكل ٣-١، نرى أن الكتلة أ مُخزَّنة في كلٍّ من عقدة البيانات ١ وعقدة البيانات ٢. ومن المهم أن تُخزَّن عدة نسخ من كل كتلة، فإن حدث خلل في إحدى عقد البيانات، يمكن لعُقدٍ أخرى أن تتولى زمام الأمور وتواصل مهام المعالجة من دون فقدٍ البيانات. لمتابعة عُقَد البيانات، إن وُجدت، وتحديد ما تعطَّل منها، تتسلم عقدة الاسم رسالةً من كلٍّ منها على حدة كلَّ ثلاث ثوانٍ، تُسمَّى «رسالة كشف أخطاء الاتصال»، وإذا لم تتسلم رسالة، فإنها تفترض أن عقدة البيانات المَعنية قد تعطَّلت عن العمل. وعليه، إذا تعطَّلت عقدة البيانات ١ في إرسال رسالة كشف أخطاء الاتصال هذه، فستصبح عقدة البيانات ٢ هي العقدة المسئولة عن عمليات الكتلة أ. ويختلف الوضع إذا فُقدت عقدة الاسم، وفي كلتا الحالتَين يجب استخدام نظام النسخ الاحتياطي المُضمَّن.



شكل ٣-١: عرضٌ مُبسَّط لجزءٍ من مجموعة في نظام هادوب للملفات الموزَّعة.

تُكتَب البيانات في عقدة البيانات لمرة واحدة، ولكنها ستُقرأ بواسطة التطبيقات لمراتٍ كثيرة. تبلغ مساحة كل كتلة عادةً ٦٤كيلوبايت فقط، ومن ثمَّ، فإن هناك الكثير منها. إحدى وظائف عقدة الاسم هي تحديد أفضل عقدة بياناتٍ لاستخدامها بناءً على الاستخدام الحالي، ما يضمن سرعة الوصول إلى البيانات ومعالجتها. ومن ثمَّ، يصل الكمبيوتر العميل إلى كتلة البيانات عبر العقدة المختارة.

تَضاف عقد البيانات طبقًا لزيادة متطلبات التخزين وعندما توجد ضرورة لذلك، وهي السمة المعروفة باسم «قابلية التوسُّع الأفقية».

إحدى المميزات الرئيسية لنظام هادوب للملفات الموزَّعة عن قواعد البيانات الارتباطية أنه يمكنك جمع كميات كبيرة من البيانات، والإضافة إليها، وذلك من دون أن تكون لديك أدنى فكرة، أثناء فعل ذلك، عمَّا تريد استخدامها من أجله. يستخدم فيسبوك، على سبيل المثال، نظام هادوب في تخزين بياناته التي تتزايد كمياتها باستمرار. والنظام لن يتسبَّب في فقد أي بيانات، كما أنه سيُخزِّن أيَّ شيءٍ وكلَّ شيءٍ في صيغته الأصلية. إنَّ إضافة عقد البيانات حسب الضرورة لا يكلِّف الكثير ولا يتطلب تغيير العُقد الموجودة بالفعل. وفي حال أصبحت العُقد التي أُنشئت سابقًا زائدةً عن الحاجة، من السهل إيقافها عن العمل. كما رأينا سابقًا، البيانات الهيكلية ذات الصفوف والأعمدة القابلة للتحديد يمكن تخزينها بسهولة في نظام إدارة قواعد بيانات ارتباطية، في حين يمكن تخزين البيانات غير الهيكلية بتكلفة منخفضة وبسهولة باستخدام أنظمة الملفات الموزَّعة.

قواعد البيانات غير الارتباطية للبيانات الضخمة

قواعد البيانات غير الارتباطية (NoSQL) هي الاسم الشامل الذي يشير إلى نوع من قواعد البيانات التي «لا تستخدم لغة الاستعلام الهيكلية فقط». لماذا طرأت الحاجة إلى نموذج غير ارتباطي لا يستخدم لغة الاستعلام الهيكلية؟ الإجابة المختصرة عن هذا السؤال هي أن النموذج غير الارتباطي يسمح لنا بإضافة بياناتٍ جديدة باستمرار. وللنموذجٍ غير الارتباطي بعض السمات الضرورية لإدارة البيانات الضخمة، وهي قابلية التوسُّع، والتوفُّر، والأداء. مع قواعد البيانات الارتباطية، لن يمكنك مواصلة التوسُّع رأسيًّا من دون خسارة قدراتها الوظيفية، بينما يمكنك التوسُّع أفقيًّا مع قواعد البيانات غير الارتباطية، الأمر الذي يسمح بالحفاظ على الأداء. قبل أن نَصِف البنية التحتية لقاعدة البيانات الموزَّعة غير الارتباطية، وسبب كونها مناسبةً للبيانات الضخمة، علينا أولًا أن نتناول نظرية CAP.

نظرية الاتساق، والتوفُّر، والسماح بخطأ انقطاع الاتصال

عام ٢٠٠٠، قدَّم إيريك بروير، أستاذ علوم الكمبيوتر في جامعة كاليفورنيا بيركلي، نظرية الاتساق والتوفُّر والسماح بخطأ انقطاع الاتصال (CAP). في سياق نظام قواعد البيانات الموزَّعة، يشير الاتساق إلى المَطلب الخاص بضرورة تماثل جميع نسخ البيانات عبر العُقد. وعليه، في شكل ٣-١، على سبيل المثال، يجب أن تكون الكتلة أ في عقدة البيانات ١ هي نفسها الكتلة أ في عقدة البيانات ٢. ويشترط التوفُّر أنه إذا تعطَّلت إحدى العُقد، فإن العُقد الأخرى تظل تؤدي وظيفتها؛ أي إذا تعطَّلت عقدة البيانات ١، فيجب أن تظل عقدة البيانات ٢ تعمل. البيانات؛ ومن ثمَّ عقد البيانات، تكون

موزّعةً فعليًّا عبر خوادم متفرقة، ويُمكن أن يتوقَف الاتصال بين هذه الأجهزة في بعض الأحيان. وعندما يحدث ذلك، فإننا نكون بصدد ما يُسمَّى بخطأ «انقطاع الاتصال في الشبكة». يتطلّب السماح بهذا الخطأ ضرورة أن يواصل النظام عمله حتى وإن حدث ذلك.

خلاصة الأمر، تنصُّ نظرية الاتساق والتوفُّر والسماح بخطأ انقطاع الاتصال على أنه فيما يخصُّ أي نظام كمبيوتر موزَّع، حيث تتم مشاركة البيانات، يمكن أن يتحقَّق معياران فقط من هذه المعايير الثلاثة. وعليه، توجد ثلاثة احتمالات، فالنظام؛ إمَّا أن يكون متسقًا ومتاحًا، وإمَّا أن يكون متسقًا ويسمح بخطأ انقطاع الاتصال في الشبكة، وإمَّا أن يسمح بخطأ انقطاع الاتصال في الشبكة ويكون متاحًا. لاحظ أنه بما أن الشبكة في نظام إدارة قواعد البيانات الارتباطية غير مُعرَّضة لخطأ انقطاع الاتصال، فإن الاتساق والتوفُّر وحدهما سيكونان المعيارَين محطَّ الاهتمام، وسيحقِّق نموذج نظام إدارة قواعد البيانات الارتباطية كلَيهما. أمَّا في قواعد البيانات غير الارتباطية، بما أن انقطاع الاتصال في الشبكة أمرٌ وارد الحدوث، فعلينا أن نختار ما بين الاتساق والتوفُّر. وإذا غضضنا الطرف عن التوفُّر، فسنتمكَّن من الانتظار حتى يتحقَّق الاتساق. أمَّا إذا اخترنا أن نغض الطرف عن الاتساق، بدلًا من ذلك، فإن هذا سيؤدي بالتبعية إلى أن البيانات ستختلف من خادم لآخر في بعض الأحيان.

توجد ثلاث خصائص تصف هذا الوضع بطريقة ملائمة وهي: متوفر دائمًا، ومرن، ومتسق في النهاية. ويبدو أن هذه الخصائص الثلاث جاءت على النقيض من الخصائص الأربع لقواعد البيانات الارتباطية. تشير كلمة «مرن» هنا إلى المرونة في متطلَّبات الاتساق. وليس الهدف هو تجاهل أيٍّ من هذه المعايير الثلاثة، بل إيجاد طريقة لتحسينها جميعها، وهي التوفيق بينها في الأساس.

بِنية قواعد البيانات غير الارتباطية

اشتُقَّت تسمية قواعد البيانات غير الارتباطية (NoSQL) من حقيقة أن لغة الاستعلام الهيكلية (SQL) لا يمكن استخدامها للاستعلام في قواعد البيانات هذه. وعليه، فإن الروابط على غرار ما رأيناه في شكل ٤ على سبيل المثال، لن تكون مُمكِنة. ثمة أربعة أنواع من قواعد البيانات غير الارتباطية: قاعدة بيانات المفتاح والقيمة، وقاعدة البيانات القائمة على الأعمدة، وقاعدة بيانات المستند، وقاعدة بيانات التمثيل البياني، وتفيد جميعها في تخزين الكميات الكبيرة من البيانات الهيكلية وشبه الهيكلية. أبسط هذه الأنواع هي قاعدة بيانات المفتاح والقيمة، وتتكوَّن من مُعرِّف (المفتاح) والبيانات المرتبطة بهذا المفتاح (القيمة)، كما هو موضَّح في جدول ٣-١. لاحظ أن «القيمة» يمكن أن تتضمَّن عناصر عديدة من البيانات.

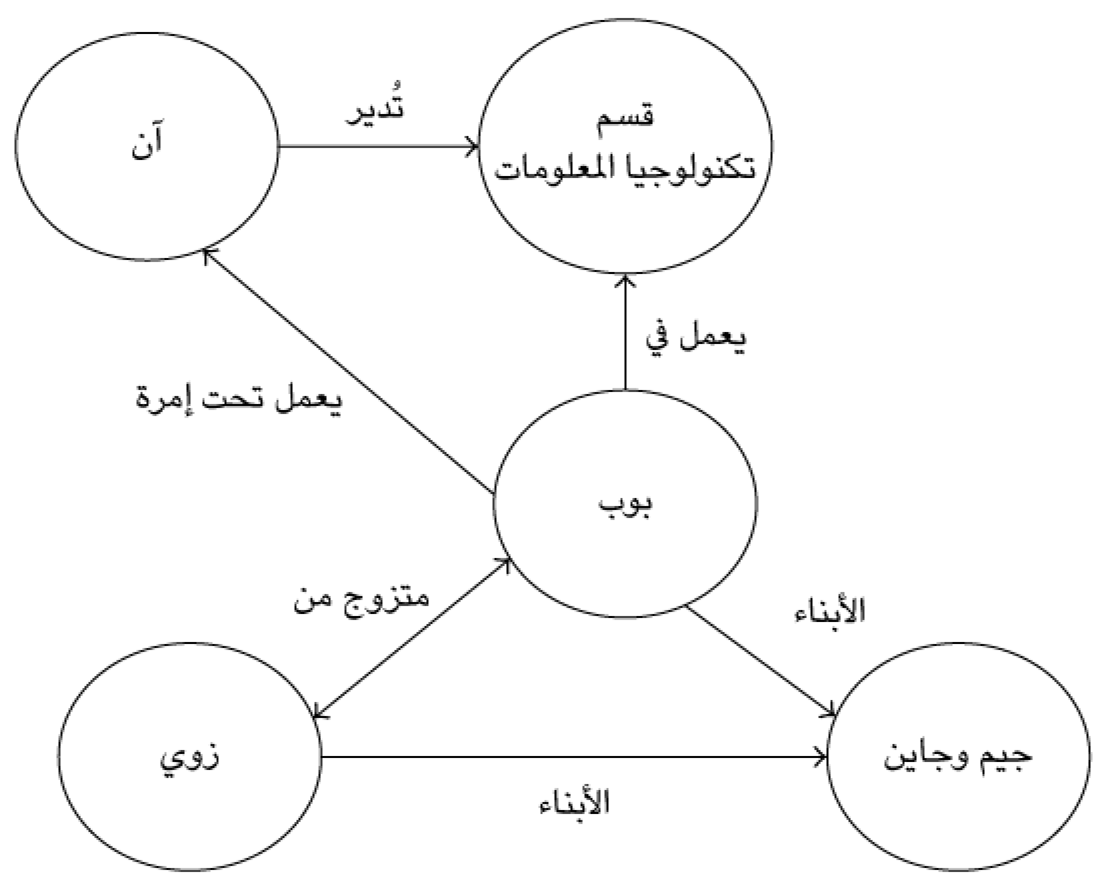
جدول ٣-١: قاعدة بيانات المفتاح والقيمة.

المفتاح	القيمة
جاين سميث	العنوان: ٣٣ أي طريق، أي مدينة
توم براون	النوع: ذكر؛ الحالة الاجتماعية: متزوج؛ عدد الأبناء: ٢؛ الأفلام المُفضَّلة: سندريلا، دراكولا، باتون

قد توجد بالطبع الكثير من أزواج المفتاح والقيمة تلك، وأن إضافة أزواج جديدة أو حذف أزواج قديمة أمر سهل للغاية؛ ممَّا يجعل قاعدة البيانات قابلةً للتوسُّع أفقيًّا إلى حدٍّ كبير. القدرة الرئيسية لهذا النوع هي أننا نستطيع البحث عن القيمة الخاصة بمفتاح معين. على سبيل المثال، باستخدام المفتاح «جاين سميث»، سنتمكَّن من العثور على عنوانها. وبتوفر كميات ضخمة من البيانات، يوفر هذا النوع من قواعد البيانات حلًّا سريعًا، وموثوقًا، وقابلًا للتوسُّع بسهولة للتخزين، ولكنه محدود بسبب عدم وجود لغة استعلام. تُعَد قواعد البيانات القائمة على الأعمدة وقواعد بيانات المستند، مُلحقَين لنموذج المفتاح والقيمة.

أمَّا قواعد بيانات التمثيل البياني، فتتبع نموذجًا مختلفًا، ويشيع استخدامها في شبكات التواصل الاجتماعي، كما تفيد في تطبيقات الأعمال. عادةً ما تكون هذه الرسوم البيانية كبيرةً للغاية، لا سيَّما عندما تُستخدَم بواسطة شبكات التواصل الاجتماعي. في هذا النوع من قواعد البيانات، تُخزَّن المعلومات في عُقَد (أو رءوس) وخطوط مستقيمة. على سبيل المثال، يوضِّح الرسم البياني في شكل ٣-٢ خمس عُقَد تصل بينها أسهم تُمثِّل العلاقات. يتغيَّر التمثيل البياني بإضافة عقدٍ أو تحديثها أو حذفها.

في هذا المثال، تُمثِّل العُقَد الأسماء والأقسام، والخطوط المستقيمة هي العلاقات بينها. وتُستخرج البيانات من التمثيل البياني عن طريق تتبُّع هذه الخطوط. إذن، إذا أردت إيجاد «أسماء موظفي قسم تكنولوجيا المعلومات الذين يعولون أطفالًا»، على سبيل المثال، فسنجد أن بوب يحقِّق هذَين المعيارَين. ولاحظ أن هذا التمثيل البياني ليس مُوجَّهًا؛ أي إننا لا نتبع اتجاهات الأسهم، بل نبحث عن وجود روابط.



شكل ٣-٢: قاعدة بيانات التمثيل البياني.

في الوقت الحالي، ثمة مقاربة تحاول الحصول على بعض الزخم تُسمَّى NewSQL. عن طريق الدمج بين أداء قواعد البيانات غير الارتباطية والخصائص الأربع للنموذج الارتباطي، فإن الهدف من هذه التقنية المُرتقَبة هو حلّ مشكلات قابلية التوسُّع المصاحبة للنموذج الارتباطي بما يجعله أكثر ملائمةً للاستخدام مع البيانات الضخمة.

التخزين السحابي

على غرار الكثير من المصطلحات الحاسوبية العصرية، يبدو مصطلح السحابة الإلكترونية مستساغًا، ومريحًا، وجذابًا، ومألوفًا، إلا أن «السحابة الإلكترونية» في الحقيقة، كما ذُكر سابقًا، مجرد طريقة للإشارة إلى شبكة من الخوادم المتصلة فيما بينها والموجودة في مراكز بيانات في جميع أنحاء العالم. وتمثِّل مراكز البيانات هذه موقعًا مركزيًّا لتخزين البيانات الضخمة.

عبر الإنترنت، نتشارك استخدام هذه الخوادم التي تَدار عن بُعد، وتوفَرها العديدُ من الشركات (بعد دفع رسوم)، في تخزين الملفات وإدارتها، وفي تشغيل التطبيقات، وما إلى ذلك. وطالما أنَّ البرنامج المطلوب لإتاحة الوصول إلى السحابة الإلكترونية موجود على الكمبيوتر أو أي جهاز آخر لديك، فسيمكنك عرض ملفاتك من أي مكان، ومنح الإذن لآخرين للوصول إليها وعرضها. كما يمكنك استخدام برنامج «موجود باستمرار» على السحابة الإلكترونية بدلًا من البرنامج الموجود على جهاز الكمبيوتر لديك. وعليه، فإن الأمر لا يتعلَّق بإمكانية الوصول إلى الإنترنت فحسب، بل يتعلَّق أيضًا بامتلاك وسيلة لتخزين المعلومات ومعالجتها، ومن هنا جاء مصطلح «الحوسبة السحابية». إن احتياجاتنا الفردية إلى التخزين السحابي ليست كبيرةً إلى هذه الدرجة، ولكن في حال زيادتها ستزيد كمية المعلومات المخزَّنة بصورة هائلة.

تُعَدُّ شركة أمازون أكبر مزوِّد للخدمات السحابية، إلا أن كمية البيانات التي تُديرها تُعامَل على أنها سر تجاري. ويمكننا أن نأخذ فكرةً عن أهمية هذه الشركة في مجال الحوسبة السحابية من خلال تناول حادثة وقعت في فبراير ٢٠١٧ عندما تعرَّضَ نظام التخزين السحابي لمنصة «خدمات أمازون ويب» (إس ثري) إلى «عطل» كبير (أي انقطاع الخدمة). استمرَّ العطل نحو خمس ساعات، ونتجَ عنه انقطاع الاتصال بالكثير من مواقع الويب والخدمات الإلكترونية، بما في ذلك نتفليكس، وإكسبيديا، وهيئة الأوراق المالية والبورصات الأمريكية. أعلنت أمازون فيما بعد أن سبب العطل كان خطأً بشريًّا، حيث ذكرت أن أحد موظفيها تسبَّب في قطع الاتصال عن الخوادم دون قصد. واستغرقت إعادة تشغيل هذه الأنظمة الضخمة وقتًا أكبر من المتوقع، ولكنها تمَّت في النهاية بنجاح. ورغم ذلك، سلَّطت هذه الحادثة الضوءَ على قابلية الإنترنت للتعطل، سواءٌ كان ذلك بسبب خطأ غير مقصود أو عملية قرصنة خبيثة المَقصِد.

ضغط البيانات غير المنقوص

في ٢٠١٧، قدَّرت مؤسسة البيانات الدولية الشهيرة أنَّ إجمالي حجم الكون الرقمي هائل ويبلغ ١زيتابايت؛ وهو عدد يتعذَّر إدراكه يساوي ١٦ × ١٠^٢١ بايت. وبالتالي، فإنه مع النمو المطرد للكون الرقمي، سيصبح من الضروري الإجابة عن الأسئلة المتعلِّقة بماهية البيانات التي يجب أن نخزِّنها فعليًّا، وعدد النسخ التي يجب الاحتفاظ بها، ومدة الاحتفاظ بها. وهذا بالأحرى يشكِّل تحديًا لوجود البيانات الضخمة؛ إذ يدفعنا إلى حذف بياناتٍ من مخازن البيانات بصورة منتظمة أو حتى أرشفتها؛ وذلك لأن هذه العملية في ذاتها مكلِّفة، ومن المحتمل أن تُفقَد بيانات قيِّمة بما أننا لا نعلم بالضرورة ماهية البيانات التي قد تكون مهمةً لنا في المستقبل. ولكن، مع كميات البيانات الهائلة الجاري تخزينها، أصبح ضغط البيانات ضروريًّا لزيادة مساحة التخزين المتاحة إلى الحد الأقصى.

ثمة تبايُن كبير في جودة البيانات التي تُجمَع إلكترونيًّا؛ ومن ثمَّ، لا بد من معالجة البيانات مسبقًا قبل تحليلها على نحو مفيد؛ وذلك من أجل الكشف عن مشكلات الاتساق والتكرار والموثوقية وعلاجها. من الواضح أن الاتساق مهم إذا كنا بصدد الاعتماد على المعلومات المستخرجة من البيانات. كما أن

إزالة التكرارات غير المرغوب فيها من تدابير الإعداد التحضيرية الجيدة لأي مجموعة بيانات، ولكن، مع وجود مجموعات البيانات الضخمة يطرأ هاجس إضافي بعدم توفر مساحة تخزين كافية للاحتفاظ بكلِّ البيانات. وعليه، تُضغَط البيانات لتقليل التكرار في مقاطع الفيديو والصور؛ ومن ثَمَّ الحد من متطلَّبات التخزين، وتحسين معدلات البث في حالة مقاطع الفيديو.

ثمة نوعان رئيسيان من الضغط: الضغط غير المنقوص والضغط المنقوص. في «الضغط غير المنقوص»، يُحتفَظ بالبيانات كلها دون فقد أي منها؛ ومن ثَمَّ، يفيد هذا النوع تحديدًا مع النصوص. على سبيل المثال، الملفات، التي لها الامتداد ZIP. تُضغَط دون فقد أي معلومات، ما يعني أن فك ضغطها يعيدنا إلى الملف الأصلي. إذا ضغطنا سلسلةً من الأحرف aaaaabbbbbbbbbb على هيئة 5a10b، فمن السهل أن نعرف كيفية فك ضغط هذه السلسلة وإعادتها مرةً أخرى إلى السلسلة الأصلية. يوجد الكثير من الخوارزميات المُستخدَمة في ضغط البيانات، ولكن سيفيدنا أولًا أن نتعرَّف على كيفية تخزين الملفات دون ضغطها.

يُعَدُّ نظام ASCII (الشفرة القياسية الأمريكية لتبادل المعلومات) الطريقة القياسية لترميز البيانات حتى يمكن تخزينها على أجهزة الكمبيوتر. يُخصَّص لكلِّ حرف أو رمز عددٌ ترتيبي، وهو رمز ASCII الخاص به. ومثلما رأينا سابقًا، تُخزَّن البيانات على هيئة سلسلة من قيم الأصفار والآحاد. يُسمَّى كلٌّ من هذه الأرقام الثنائية «بِت». ويستخدِم نظام ASCII القياسي ٨بت (وهو ما يُعرَف أيضًا بأنه يعادل ١بايت) لتخزين كل حرف ورمز. على سبيل المثال، في نظام ASCII، يُرمَز للحرف a بالعدد ٩٧ والذي يتحوَّل إلى ٠١١٠٠٠٠١ بالنظام الثنائي. هذه القيم مُدرجَة في جدول نظام ASCII القياسي، الذي وضعنا جزءًا صغيرًا منه في نهاية هذا الكتاب. وفيما يخصُّ الأحرف الإنجليزية الكبيرة، فإنَّ لها رموزًا مختلفة في نظام ASCII.

دعونا نتناول كيفية ترميز سلسلة الأحرف added كما هو معروضٌ في جدول ٣-٢.

جدول ٣-٢: سلسلة أحرف بعد ترميزها

سلسلة الأحرف	d	e	d	d	a
ASCII	١٠٠	١٠١	١٠٠	١٠٠	٩٧
النظام الثنائي	٠١١٠٠١٠٠	٠١١٠٠١٠١	٠١١٠٠١٠٠	٠١١٠٠١٠٠	٠٠٠٠١

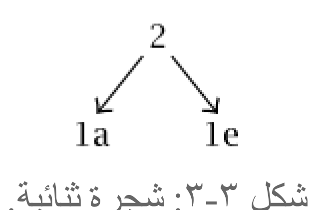
إذن، تَشْغَل سلسلة الأحرف added مساحة تخزين مقدارها ٥بايت أو ٥ × ٨ = ٤٠بت. وبالنظر إلى شكل ٧، يتحقَّق فك الترميز باستخدام جدول رموز ASCII. لا تُعَد هذه طريقةً اقتصادية لترميز

البيانات وتخزينها؛ إذ إن تعيين ٨بت لكل حرف يبدو مساحةً مُبالغًا فيها ولا تأخذ في الاعتبار حقيقة أن بعض الحروف في المستندات النصية تُستخدَم بوتيرة أكثر تكرارًا من غيرها. يوجد الكثير من نماذج ضغط البيانات دون فقدها، مثل خوارزمية هوفمان، التي تستخدم مساحة تخزين أقل عن طريق الترميز المتغير الطول، وهو أسلوب يعتمد على مدى تكرار حرف معين. تُعيَّن للأحرف الأكثر تكرارًا رموز أقصر طولًا.

بالرجوع إلى سلسلة الأحرف added مرةً أخرى، نلاحظ أن الحرف a يظهر مرةً واحدة، والحرف e يظهر مرةً واحدة، والحرف d يظهر ثلاث مرات. وبما أن الحرف d هو الأكثر تكرارًا، فلا بد أن يُخصَّص له أقصر رمز. لإيجاد رمز هوفمان لكل حرف، نَعِد الأحرف المكوِّنة لكلمة added على النحو الآتي:

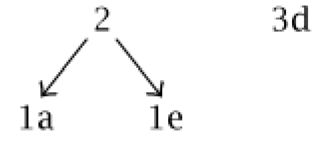
$$1a \longrightarrow 1e \longrightarrow 3d$$

بعد ذلك، نبحث عن الحرفَين الأقل تكرارًا؛ أي الحرفَين a وe، ثم نُنشئ التركيب الموضَّح في شكل ٣-٣، ويُسمَّى «الشجرة الثنائية». العدد ٢ في أعلى الشجرة هو حاصل جمع عدد مرات تكرار الحرفَين الأقل تكرارًا.



شكل ٣-٣: شجرة ثنائية.

نوضِّح في شكل ٣-٤ العقدة الجديدة التي تمثِّل التكرارات الثلاثة للحرف d.



شكل ٣-٤: الشجرة الثنائية مع إضافة عقدة جديدة.

يوضِّح شكل ٣-٤ الشجرة الكاملة وقد وُضِع في أعلاها العددُ الإجمالي لتكرارات الحرف. ويتم ترميز كل طرفٍ في الشجرة إمَّا بصفر وإمَّا بواحد، كما هو موضَّح في شكل ٣-٥، ويكون التوصل إلى الرموز المقابلة من خلال تتبُّع المسارات حتى أعلى الشجرة.

5
0 1
2 3d
0 1
1a 1e

الحرف	الرمز (بالبت)
a	٠٠
e	١٠
d	١

شكل ٣-٥: الشجرة الثنائية كاملة.

وعليه، يتم ترميز كلمة added كالآتي: a = ٠٠، وd = ١، وd = ١، وe = ١٠، وd = ١، وهذا يعطينا ٠٠١١١٠١. باستخدام هذه الطريقة نلاحظ أن مساحة التخزين المستخدمة هي ٣بت لتخزين الحرف d، و٢بت لتخزين الحرف a، و٢بت لتخزين الحرف e، ما يعطينا مساحةً إجمالية مقدارها ٧ بت. وهذه المساحة أفضل بكثير من المساحة الأصلية التي تبلغ ٤٠بت.

ثمة طريقة لقياس مدى كفاءة عملية الضغط، وهي حساب نسبة ضغط البيانات، وتُعرَّف بأنها حجم أحد الملفات دون ضغط مقسومًا على حجمه مضغوطًا. في هذا المثال، النسبة ٧/٤٥ تساوي تقريبًا ٦٫٤٣، وهي نسبة ضغط عالية تدل على توفير جيد لمساحة التخزين. ومن الناحية العملية، تكون هذه الأشجار كبيرةً للغاية؛ ومن ثمَّ تُستخدَم أساليب رياضية معقَّدة لتحسينها. وهكذا يوضِّح لنا هذا المثال كيف يمكننا ضغط البيانات دون فقد أيٍّ من المعلومات المتضمَّنة في الملف الأصلي، ومن هنا جاءت تسمية هذا النوع من ضغط البيانات بالضغط غير المنقوص.

ضغط البيانات المنقوص

في المقابل، عادةً ما تكون ملفات الصوت والصور أكبر بكثير من ملفات النصوص؛ ومن ثمَّ، يُستخدم معها أسلوب مختلف يُسمَّى «الضغط المنقوص». ويرجع هذا إلى أنَّ تطبيق أساليب الضغط غير المنقوص عند التعامل مع ملفات الصوت والصور قد لا يُسفر عن نسبة ضغط عالية بما يكفي ليكون تخزين البيانات بهذه الطريقة مُجديًا. هذا بالإضافة إلى أن فقد بعض البيانات من ملفات الصوت والصور أمرٌ مقبول. يستغل الضغط المنقوص هذه السمة الأخيرة، ويحذف بعض البيانات في الملف الأصلي؛ ومن ثمَّ يُقلِّل من مساحة التخزين اللازمة. تتمحور الفكرة الرئيسية حول حذف بعض التفاصيل دون أن يؤثِّر ذلك بدرجة كبيرة على إدراكنا للصورة أو الصوت.

على سبيل المثال، لنفترض أنَّ لدينا صورةً فوتوغرافية بالأبيض والأسود، أو بوصفٍ أدق «صورة ذات تدرُّج رمادي»، لطفل يتناول الآيس كريم على شاطئ البحر. يحذف الضغط المنقوص كميتَين متماثلتَين من البيانات من صورة الطفل ومن صورة البحر. تُحسَب نسبة البيانات المحذوفة بحيث لا

يكون لها تأثير كبير على إدراك الناظِر للصورة الناتجة (المضغوطة)؛ فالضغط المفرط سيؤدي إلى صورة مُشوَّشة. ذلك حيث تأتي زيادة مستوى الضغط على حساب جودة الصورة.

إذا أردنا ضغط صورة ذات تدرُّج رمادي، فإننا نقسِّمها أولًا إلى مربعات تبلغ مساحة كلٍّ منها ٨ بكسل × ٨ بكسل. وبما أن هذه المساحة صغيرة للغاية، فستكون جميع وحدات البكسل متشابهةً بوجه عام من حيث الشكل. ومن ثمَّ، تمثِّل هذه الملاحظة، بالإضافة إلى الإلمام بآلية إدراكنا للصور، أحد أساسيات الضغط المنقوص. يحتوي كلُّ بكسل على قيمة عددية تتراوح ما بين صفر للأسود الخالص و٢٥٥ للأبيض الخالص، وتُمثِّل الأعدادُ التي تتدرج بينهما ظلال اللون الرمادي. وبعد إجراء بعض المعالجة الإضافية باستخدام طريقة تُسمَّى «خوارزمية جيب التمام المتقطع»، نحصل على متوسط قيمة الكثافة لكل كتلة، وتُقارَن النتائج مع كلٍّ من القيم الفعلية لكتلةٍ معينة. وبما أننا نقارن هذه القيم الفعلية بمتوسط القيمة، فإن معظمها سيكون صفرًا أو سيصبح صفرًا عند تقريبه. تجمع خوارزمية الضغط المنقوص جميع هذه الأصفار معًا، وهو ما يمثِّل المعلومات المأخوذة من وحدات البكسل الأقل أهميةً بالنسبة إلى الصورة. تُجمَع كلٌّ هذه القيم، التي تناظر المناطق ذات الترددات العالية في الصورة، معًا وتُحذَف المعلومات المُكرَّرة، باستخدام أسلوبٍ يُسمَّى «التكميم»؛ ومن ثمَّ يحدث الضغط. على سبيل المثال، إذا كان لدينا ٦٤ قيمةً يلزم لتخزين كلٍّ منها بايت واحد، وكان لدينا ٢٠ صفرًا، فإن كلَّ ما سنحتاجه بعد الضغط هو مساحة تخزين مقدارها ٤٥بايت فقط. وتُكرَّر هذه العملية مع جميع الكتل المكوِّنة للصورة؛ ومن ثمَّ تُحذَف المعلومات المُكرَّرة منها جميعًا.

فيما يخصُّ الصور الملوَّنة، تتعرَّف خوارزمية «جيه بي إي جي» (المجموعة المشتركة لخبراءِ التصوير الفوتوغرافي)، على سبيل المثال، على الألوان الأحمر والأزرق والأخضر، وتُعيِّن لكلٍّ منها بُعدًا مختلفًا بِناءً على الخصائص المعروفة للإدراك البصري لدى البشر. يُعيَّن لِلَّون الأخضر أقصى بُعد؛ لأن العين البشرية أكثر إدراكًا للَّون الأخضر عن اللَّونَين الأحمر والأزرق. ويُعيَّن لكل بكسل في الصور الملوَّنة قيمة كثافة لمكونات اللون الأحمر والأزرق والأخضر فيها، ويُمثَّل هذا بالقيمة الثلاثية R, G, B. ولأسباب تِقْنيَّة، عادةً ما تُحوَّل قيم R, G, B الثلاثية إلى قيمة ثلاثية أخرى، مثل YCbCr حيث يمثِّل حرف Y كثافة اللون، وكلٌّ من Cb وCr هما قيمتا التشبُّع اللوني، اللتان تصفان اللون الفعلي. وباستخدام خوارزمية رياضية معقَّدة، يمكن تقليل قيم كل بكسل وإجراء ضغط منقوص في نهاية المطاف من خلال تقليل عدد وحدات البكسل المحفوظة.

بوجه عام، يتم ضغط ملفات الوسائط المتعدِّدة باستخدام أساليب الضغط المنقوص نظرًا لأحجامها الكبيرة. فكلما زاد مستوى ضغط الملف، تراجعت جودة إعادة إنتاجه، ولكن، نظرًا للتضحية ببعض البيانات، يمكن تحقيق نسب ضغط أكبر، بما يجعل الملف أصغر حجمًا.

بعد أن وضعت المجموعةُ المشتركة لخبراء التصور الفوتوغرافي معيارًا دوليًّا لضغط الصور للمرة الأولى عام ١٩٩٢، أصبحَ تنسيق الملف JPEG هو الطريقة الأكثر شيوعًا لضغط الصور الفوتوغرافية سواءٌ الملونة أم ذات التدرج الرمادي. ولا تزال هذه المجموعة تزاول نشاطها وتعقد اجتماعاتها عدة مرات كلَّ عام.

لنَعُد مجدَّدًا إلى مثال الصورة الفوتوغرافية ذات اللونَين الأبيض والأسود الملتقطة لطفل يتناول الآيس كريم على شاطئ البحر. من البديهي أن نراعي عند ضغط هذه الصورة أن يظل الجزء الذي يُظهر الطفل واضحًا؛ ومن ثمَّ فإننا نُضحِّي في سبيل ذلك بجزءٍ من وضوح تفاصيل الخلفية. أصبح هذا الأمر ممكنًا مع الطريقة الجديدة المسماة «ضغط البيانات عن طريق تمويهها»، التي ابتكرها الباحثون في كلية هنري ساموييلي للهندسة والعلوم التطبيقية، بجامعة كاليفورنيا في لوس أنجلوس. وفيما يخصُّ القراء المهتمين بالتفاصيل، يُرجى مراجعة جزء «قراءات إضافية» في نهاية هذا الكتاب.

رأينا كيف يمكن استخدام نظام ملفات البيانات الموزَّعة في تخزين البيانات الضخمة. وتمَّ التغلُّب على مشكلات التخزين، حتى إنه أصبح من الممكن حاليًّا استخدام مصادر البيانات الضخمة في الإجابة عن الأسئلة التي كانت تتعذَّر الإجابة عنها سابقًا. وكما سنرى في الفصل الرابع، تُستخدم طريقة خوارزمية تُسمَّى «ماب رديوس» في معالجة البيانات المخزنة في نظام هادوب للملفات الموزَّعة.

الفصل الرابع

تحليلات البيانات الضخمة

بعدما تحدَّثنا عن كيفية جمع البيانات الضخمة وتخزينها، يمكننا الآن تناول بعضٍ من الأساليب المُستخدمة في استخراج المعلومات المفيدة من هذه البيانات، على غرار تفضيلات العملاء أو مدى سرعة انتشار وباءٍ ما. تتغيَّر تحليلات البيانات الضخمة، المصطلح الشامل لأساليبِ تحليل البيانات، بسرعة مع تزايد أحجام مجموعات البيانات وإفساح علم الإحصاء التقليدي المجالَ أمام هذا النموذج الجديد.

تقدِّم شركة هادوب، التي تعرَّفنا عليها في الفصل الثالث، وسيلةً لتخزين البيانات الضخمة من خلال نظام الملفات الموزَّعة الذي أنشأته. وكمثال على أساليب تحليل البيانات الضخمة، سنلقي نظرةً على نموذج «ماب رديوس» البرمجي، وهو عبارة عن نظام معالجة للبيانات الموزَّعة والذي يشكِّل جزءًا من الوظيفة الأساسية لنظام «هادوب إيكوسيستم». تستخدم أمازون، وجوجل، وفيسبوك، وغيرها من مؤسساتٍ أخرى نظام هادوب في تخزين بياناتها ومعالجتها.

نموذج «ماب رديوس»

إحدى الطرق الشائعة للتعامل مع البيانات الضخمة هي تقسيمها إلى مجموعاتٍ صغيرة ثم معالجة كلٍّ من هذه المجموعات على حدة، وهذا ما يفعله نموذج «ماب رديوس» MapReduce في الأساس عن طريق توزيع العمليات الحسابية أوِ الاستعلامات المطلوبة على الكثير والكثير من أجهزة الكمبيوتر. من المفيد أن نتناول مثالًا مبسَّطًا ومختصرًا على آلية عمل «ماب رديوس»، وبما أننا سنفعل ذلك يدويًّا، لا بد أن يكون حقًّا مثالًا مختصرًا إلى حدٍّ كبير، ولكنه يوضِّح في الوقت نفسه الآلية المستخدمة مع البيانات الضخمة. لا توجد بطبيعة الحال عدة آلاف من المعالجات المستخدَمة في معالجة كمية ضخمة من البيانات على نحو متزامن، ولكن، هذه الآلية قابلة للتوسُّع، وهي فكرةٌ بارعة ومن السهل تطبيقها.

ثمة أجزاء عديدة في هذا النموذج التحليلي: مُكَوِّن «التجزئة»، ومرحلة «الخلط»، ومُكَوِّن «التجميع». يُكتَب مُكَوِّن التجزئة بواسطة المُستخدِم، ويجري فيه فرز البيانات التي تُهمنا. ثم تأتي مرحلة الخلط، التي تُعَد جزءًا رئيسيًّا من الكود الرئيسي لنموذج «ماب رديوس» من هادوب، حيث تُوضَع البيانات في مجموعاتٍ حسب المفتاح، وأخيرًا نصل إلى مُكَوِّن التجميع، الذي يكتبه المستخدم أيضًا، والذي يتولَّى تجميع هذه المجموعات وعرض النتيجة. تُرسَل النتيجة بعد ذلك إلى نظام هادوب للملفات الموزَّعة من أجل تخزينها.

لنفترض، على سبيل المثال، أن لدينا ملفات المفتاح والقيمة الآتية مُخزّنة في نظام هادوب للملفات الموزَّعة، مع وجود إحصاءاتٍ عن كلٍّ من الأمراض التالية: الحصبة، وفيروس زيكا، والسُّل، والإيبولا. في هذه الحالة، المرض هو المفتاح، وتُعيَّن له قيمة تمثِّل عدد الحالات المصابة بكل مرض. ما يهمنا هو إجمالي عدد الحالات لكل مرض.

الملف ١:

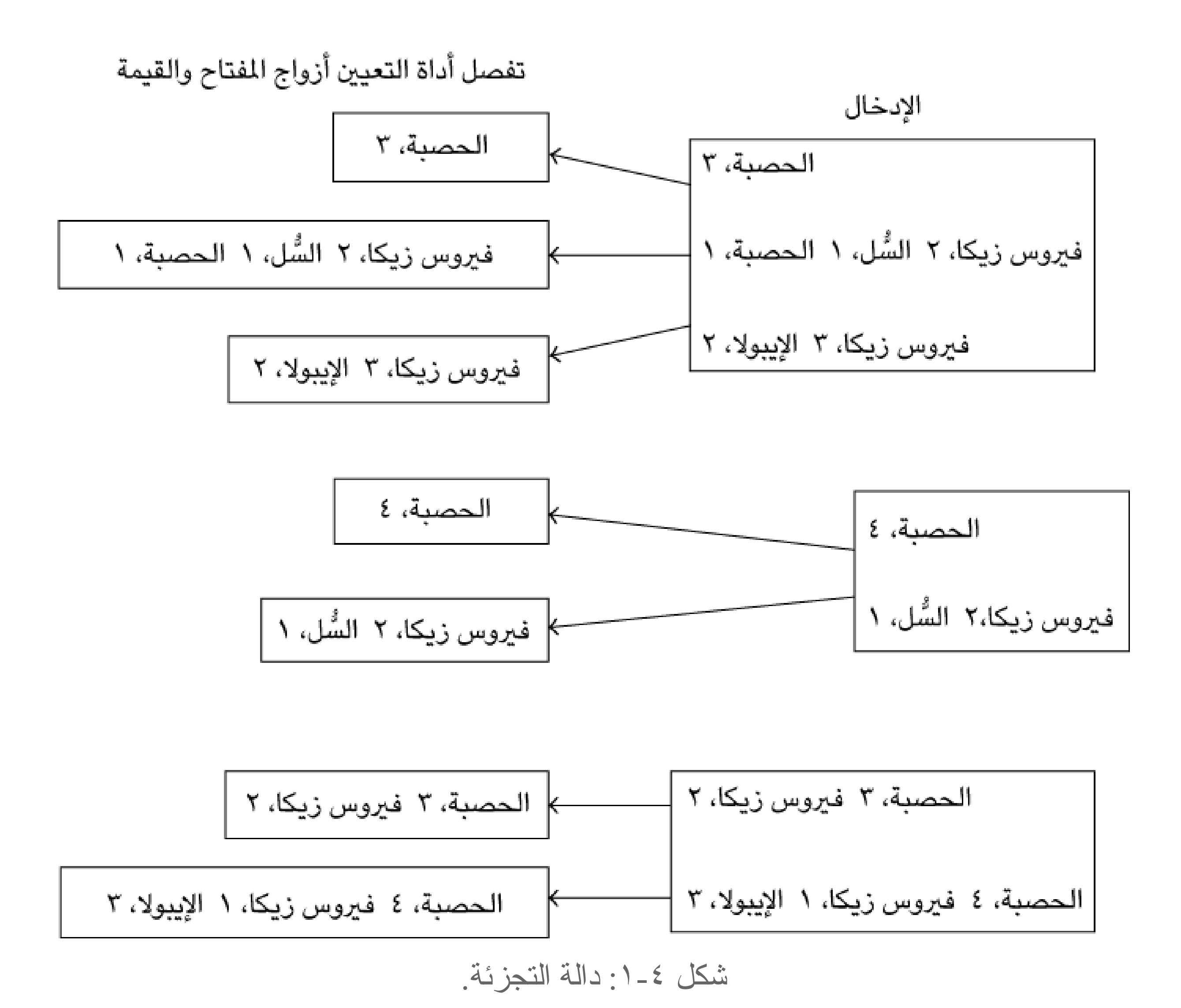
الحصبة، ٣.
فيروس زيكا، ٢ السُّل، ١ الحصبة، ١.
فيروس زيكا، ٣ الإيبولا، ٢.

الملف ٢:

الحصبة، ٤.
فيروس زيكا، ٢ السُّل، ١.

الملف ٣:

الحصبة، ٣ فيروس زيكا، ٢.
الحصبة، ٣ فيروس زيكا، ٢.
الحصبة، ٤ فيروس زيكا، ١ الإيبولا، ٣.

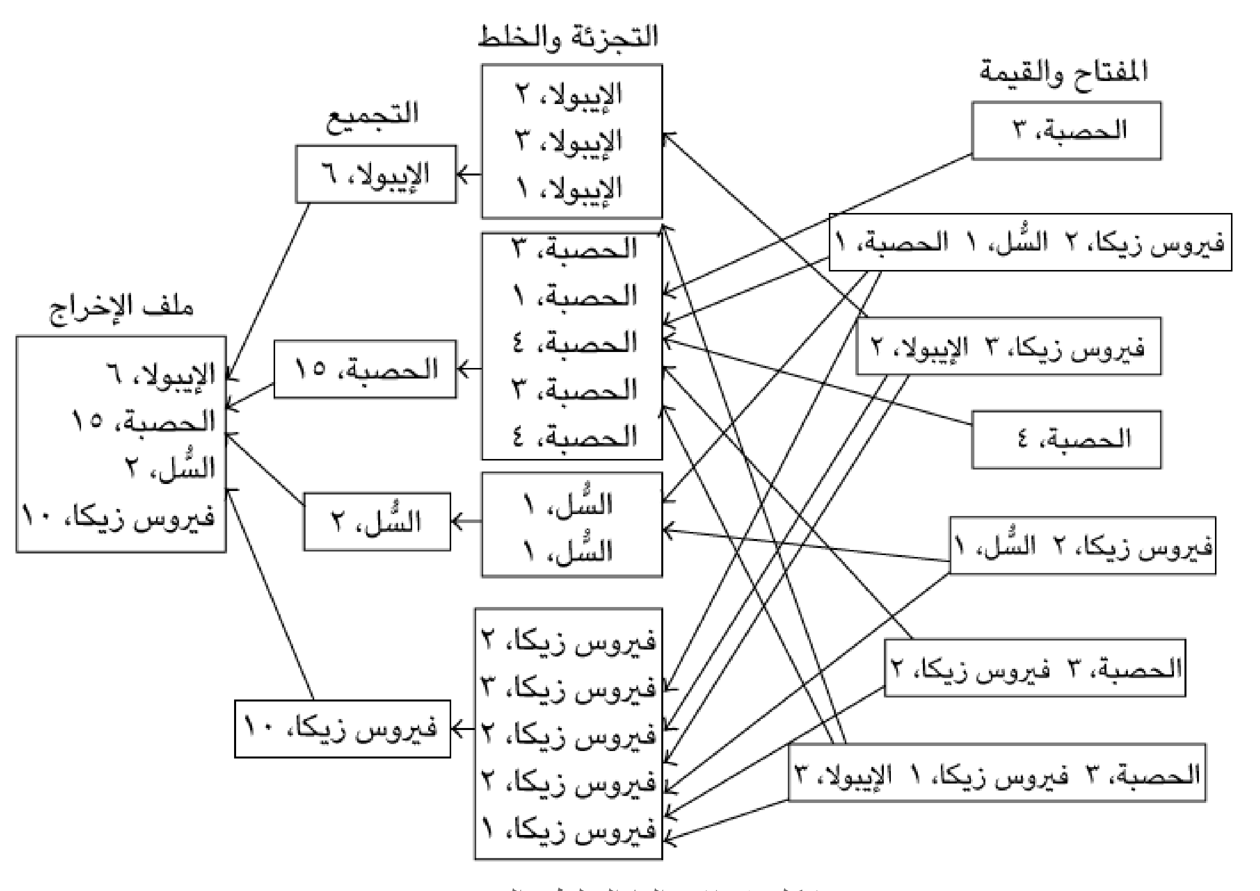


شكل ٤-١: دالة التجزئة.

تمكِّننا أداة التعيين من قراءة كل ملف من ملفات الإدخال هذه على حدة، سطرًا بسطر، كما هو موضَّح في شكل ٤-١. ثم تعرض أداة التعيين نتيجةً بأزواج المفتاح والقيمة لكلٍّ من هذه السطور المنفردة.

بعد تجزئة الملفات وإيجاد أزواج المفتاح والقيمة لكل ملف مجزَّأ، تُستخدَم في الخطوة التالية الخوارزمية التي يقدِّمها البرنامج الرئيسي، والتي تتولَّى فرز أزواج المفتاح والقيمة وخلطها. يُجرى فرزٌ أبجدي للأمراض، وتُرسَل النتيجة إلى ملف مناسب استعدادًا لعملية التجميع، كما هو موضَّح في شكل ٤-٢.

استمرارًا مع شكل ٤-٢، يدمج مُكوِّن التجميع نتائج مرحلتَي التجزئة والخلط، ونتيجةً لذلك، يرسل بيانات كل مرض إلى ملف منفصل. بعد ذلك، تسمح مرحلة التجميع في الخوارزمية بحساب الإجماليات الفردية ثم تُرسِل هذه النتائج إلى ملف إخراج نهائي، في صورة أزواج المفتاح والقيمة، يمكن حفظه في نظام الملفات المُوزَّعة.



شكل ٤-٢: دالتا الخلط والتجميع.

يُعَد هذا مثالًا بسيطًا للغاية، ولكن يُمَكِّننا نموذج «ماب رديوس» من تحليل كمياتٍ كبيرةٍ للغاية من البيانات. على سبيل المثال، باستخدام البيانات التي تقدِّمها مؤسسة كومون كراول، وهي مؤسسة غير ربحية توفِّر نسخةً مجانية من شبكة الإنترنت، يمكننا إحصاء عدد مرات ظهور كل كلمة على شبكة الإنترنت عن طريق كتابة برنامج كمبيوتر مناسب يستخدم نموذج «ماب رديوس» البرمجي.

عوامل تصفية «بلوم»

أحد الأساليب المفيدة بوجه خاص في التنقيب في البيانات الضخمة عامل تصفية «بلوم» Bloom، وهو أسلوبٌ يعتمد على نظرية الاحتمال ابتُكِرَ في سبعينيات القرن العشرين. كما سنرى، تناسب عوامل تصفية «بلوم» بشكل خاص التطبيقات التي يُمثِّل فيها التخزين مشكلة، والتي يمكن فيها التفكير في البيانات على هيئة قائمة.

الفكرة الأساسية في عوامل تصفية «بلوم» أننا نريد إنشاء نظام، بِناءً على قائمة من عناصر البيانات، للإجابة عن السؤال «هل يوجد (س) في القائمة؟» في حالة مجموعات البيانات الضخمة،

قد تصبح عملية البحث في المجموعة بأكملها بطيئةً للغاية بما يجعل البحث غير مفيد؛ ومن ثمَّ، نستخدم عامل تصفية «بلوم» الذي لا يكون دقيقًا بنسبة ١٠٠ في المائة؛ لأنه أسلوب قائم على الاحتمالات — قد تحدِّد الخوارزمية أن عنصرًا ما ينتمي إلى القائمة رغم أنه لا ينتمي إليها في الواقع، ولكنه مع ذلك أسلوبٌ يتسم بالسرعة والموثوقية وكفاءة التخزين لاستخراج المعلومات المفيدة من البيانات.

لعوامل تصفية «بلوم» الكثير من الاستخدامات. على سبيل المثال، يمكن استخدامها للتحقق ممَّا إذا كان عنوان ويب معين يؤدي إلى موقع إلكتروني ضار. في هذه الحالة، قد يعمل عامل تصفية «بلوم» بمثابة قائمة حظر بعناوين المواقع الإلكترونية الضارة المعروفة والتي يمكن الاستعانة بها للتحقُّق، بسرعة ودقة، ممَّا إذا كان عنوان الموقع الإلكتروني الذي نقرت عليه للتوِّ آمنًا أم لا. ويمكن كذلك أن نضيف إلى قائمة الحظر عناوين المواقع الإلكترونية التي اكتُشِفَ حديثًا أنها ضارة. وبما أن عدد المواقع الإلكترونية قد تخطَّى حاليًّا المليار، ويُضاف المزيد منها يوميًّا، أصبح تتبُّع المواقع الإلكترونية الضارة إحدى مشكلات البيانات الضخمة.

أحد الأمثلة الوثيقة الصلة على ذلك رسائل البريد الإلكتروني الضارة، التي ربما تكون بريدًا عشوائيًّا أو ربما تحتوي على محاولات تصيُّد احتيالي. يوفِّر لنا عامل تصفية «بلوم» طريقة سريعة للتحقق من كل عنوان بريد إلكتروني؛ ومن ثمَّ، نتمكَّن من إصدارِ تحذير في الوقت المناسب إذا لزم الأمر. يشغل كل عنوان ٢٠بايت تقريبًا؛ ومن ثمَّ فإنَّ تخزين كلٍّ منها وفحصه عملية تستهلك وقتًا طويلًا للغاية بما أننا نحتاج إلى تنفيذ هذا بسرعة كبيرة — باستخدام عامل تصفية «بلوم»، يمكننا تقليل كمية البيانات المخزَّنة كثيرًا. يمكننا أن نرى كيفية هذا عمليًّا باتباع عملية إنشاء عامل تصفية «بلوم» صغير واستعراض آلية عمله.

لنفترض أن لدينا قائمةً بعناوين البريد الإلكتروني التالية، ونريد أن نضع عليها علامةً بأنها ضارة: aaa@aaaa.com; bbb@nnnn.com; ccc@ff.com; dd@ggg.com. لإنشاء عامل تصفية «بلوم» الخاص بنا، نفترض أولًا أن مساحة الذاكرة المتوافرة لدينا على جهاز الكمبيوتر هي ١٠بت. يُطلَق على هذا اسم «مصفوفة البت»، وتكون فارغةً في البداية. للبت حالتان فقط، يُرمز لهما عادةً بصفر وواحد؛ ولذا سنبدأ بضبط كل القيم في مصفوفة البت على ٠، ما يعني أنها فارغة. وكما سنرى بعد قليل، سيعني البت الذي قيمته ١ أن الفهرس المرتبط قد تمَّ تعيينه لمرة واحدة على الأقل.

حجم مصفوفة البت ثابت، وسيظل كما هو دون تغيير بغض النظر عن عدد الحالات التي نُضيفها. وننشئ فهرسًا لكل بت في المصفوفة كما هو موضَّح في جدول ٤-١.

جدول ٤-١: مصفوفة مكوَّنة من ١٠بت

الفهرس	٠	١	٢	٣	٤	٥	٦	٧	٨	٩
قيمة البت	٠	٠	٠	٠	٠	٠	٠	٠	٠	٠

علينا الآن أن نتعرَّف على «دوال التجزئة»، وهي عبارة عن خوارزميات مُصمَّمة لتعيين كل عنصر في قائمة معينة إلى موضع ما في المصفوفة. وبذلك، لن نخزِّن سوى الموضع المُعيَّن في المصفوفة، بدلًا من عنوان البريد الإلكتروني نفسه؛ ومن ثمَّ يقل مقدار مساحة التخزين المطلوبة.

في شرحنا هنا، سنعرض نتيجة استخدام دالتَي تجزئة، ولكن، تُستخدم في المعتاد ١٧ أو ١٨ دالةً معًا في حالة التعامل مع مصفوفة أكبر بكثير. وبما أن هذه الدوال مُصمَّمَة لإجراء عملية التعيين على نحو موحَّد نوعًا ما، فإن كلَّ فهرس لديه فرصة متساوية لعرضه كنتيجة في كل مرة تُطبَّق فيها خوارزمية التجزئة على عنوان مختلف.

ومن ثمَّ، علينا أولًا أن ندع خوارزميات التجزئة تُعيِّن كلَّ عنوان بريد إلكتروني إلى أحد فهارس المصفوفة.

لإضافة العنوان الإلكتروني aaa@aaaa.com إلى المصفوفة، يُمرَّر أولًا عبر دالة التجزئة ١، التي تعرض قيمة موضع أو فهرس داخل المصفوفة. على سبيل المثال، دعونا نفترض أن دالة التجزئة ١ عرضت الفهرس ٣. وعند تطبيق دالة التجزئة ٢ على العنوان الإلكتروني aaa@aaaa.com، عرضت الفهرس ٢. سيكون لكلٍّ من هذَين الموضعَين قيمة بت مُخزَّنة لهما مُعيَّنة على ١. إذا كان الموضع مُعيَّنًا على القيمة ١ بالفعل، فإنه يُترَك كما هو. وبالمثل، قد ينتج عن إضافة العنوان الإلكتروني bbb@nnnn.com في الموضعَين ٢ و٧ شغل هذَين الموضعَين أو تعيين القيمة ١ لهما، وقد ينتج عن إضافة العنوان الإلكتروني ccc@ff.com الموضعان ٤ و٧. وأخيرًا، افترض أن دالتَي التجزئة المُطبقتَين على العنوان الإلكتروني dd@ggg.com ينتج عنهما الموضعَين ٢ و٦. يعرض جدول ٤-٢ ملخصًا بهذه النتائج.

جدول ٤-٢: ملخص نتائج دالتَي التجزئة

البيانات	دالة التجزئة ١	دالة التجزئة ٢
aaa@aaa.com	٣	٤
bbb@nnnn.com	٢	٧
ccc@ff.com	٤	٧
dd@ggg.com	٢	٦

مصفوفة عامل تصفية «بلوم» الحقيقي موضَّحة في جدول ٤-٣ مع تعيين القيمة ١ إلى المواضع المشغولة فيها.

جدول ٤-٣: عامل تصفية «بلوم» لعناوين البريد الإلكتروني الضارة

الفهرس	٠	١	٢	٣	٤	٥	٦	٧	٨	٩
قيمة البت	٠	٠	١	١	١	٠	١	١	٠	٠

كيف نستخدم إذن هذه المصفوفة بوصفها عامل تصفية «بلوم»؟ دعونا نفترض أننا تسلمنا رسالة بريد إلكتروني ونرغب في التحقق ممَّا إذا كان العنوان يظهر في قائمة عناوين البريد الإلكتروني الضارة أم لا. ولنفترض أن هذا العنوان مُعيَّن إلى الموضعَين ٢ و٧، اللذَين يحملان القيمة ١. بما أن جميع القيم المعروضة تساوي ١، من «المحتمل» أن العنوان ينتمي إلى القائمة، وعليه، فمن «المحتمل» أن يكون ضارًّا. لا يمكننا الجزم يقينًا بأن العنوان موجود في القائمة؛ لأن الموضعَين ٢ و٧ كانا نتيجة تعيين عناوين أخرى وربما تكون الفهارس قد استُخدِمَت أكثر من مرة. ومن ثمَّ، عندما نختبر انتماء عنصر ما إلى القائمة، فمن المحتمَل أيضًا عرضُ نتيجة إيجابية خاطئة. ومع ذلك، في حالة عرض فهرس مصفوفة بالقيمة ٠ كنتيجة لأي دالة تجزئة (تذكَّر أنه قد يوجد بوجه عام ١٧ أو ١٨ دالة)، نعلم يقينًا أن العنوان غير موجود في القائمة.

إن العمليات الحسابية المتضمَّنة معقَّدة، ولكننا نلاحظ أنه كلما زاد حجم المصفوفة زاد عدد الأماكن غير المشغولة، وتضاءلت احتمالية الحصول على نتائج إيجابية زائفة أو تطابق غير صحيح. ومن الجليِّ أن حجم المصفوفة يتحدَّد بعدد المفاتيح ودوال التجزئة المُستخدَمة، ولكن لا بد أن تكون المصفوفة كبيرةً بما يكفي لتوفير عدد من الأماكن غير المشغولة يسمح لعامل التصفية بأداء وظيفته بفاعلية ويقلِّل عدد النتائج الإيجابية الزائفة إلى الحد الأدنى.

تتسم عوامل تصفية «بلوم» بالسرعة، ويمكنها أن تقدِّم طريقةً مفيدة للغاية لاكتشاف مُعامَلات بطاقات الائتمان الاحتيالية. يتحقَّق عاملُ التصفية ممَّا إذا كان عنصرٌ معين ينتمي إلى قائمة أو مجموعة معينة أم لا، وعليه، تُوضَع علاماتٌ على أي معاملات غير معتادة بأنها لا تنتمي إلى قائمة معاملاتك المعتادة. على سبيل المثال، إذا لم تكن اشتريت من قبل معدات تسلق الجبال باستخدام بطاقتك الائتمانية، فسيضع عامل تصفية «بلوم» علامةً على عملية شراء حِبَال التسلق هذه بأنها مشبوهة. وعلى النقيض من ذلك، إذا كنت قد اشتريت معدات تسلق الجبال من قبل، فسيُحدِّد عامل تصفية «بلوم» عملية الشراء هذه بأنها ربما تكون مقبولة، ولكن سيظل الاحتمالُ قائمًا في أن النتيجة زائفة.

كما يمكن استخدام عوامل تصفية «بلوم» لتصفية رسائل البريد الإلكتروني بحثًا عن البريد العشوائي. وتُعد عوامل تصفية البريد العشوائي مثالًا جيدًا على ذلك بما أننا لا نعلم ما نبحث عنه بالضبط — نحن نبحث عادةً عن أنماط، ومن ثمَّ، إذا كنا نريد أن تُعامَل رسائل البريد الإلكتروني المشتملة على كلمة mouse على أنها بريد عشوائي، فعلينا أن نحدِّد أيضًا ضرورة التعامل مع أشكال أخرى للكلمة، مثل m0use أو mou$e، على أنها بريد عشوائي. في الواقع، نريد التعامل مع كل الأشكال المحتمَلة التي يمكن تعريفها للكلمة على أنها بريد عشوائي. سيكون من الأسهل كثيرًا تصفية كل الكلمات التي لا تطابق كلمةً معينة، وعليه، قد نسمح لكلمة mouse وحدها بالمرور عبر عامل التصفية.

تَستخدَم عوامل تصفية «بلوم» أيضًا لتسريع الخوارزميات المُستخدَمة في ترتيب نتائج استعلامات الويب، وهو موضوع على جانب كبير من الأهمية لأولئك الذين لديهم مواقع إلكترونية يرغبون في الترويج لها.

خوارزمية «بيدج رانك»

عندما نُجري بحثًا على محرك بحث جوجل، تُرتَّب المواقع الإلكترونية الناتجة حسب صلتها بكلمات البحث. يجري محرك بحث جوجل هذا الترتيب في الأساس عن طريق تطبيق خوارزمية تُسمَّى «بيدج رانك» PageRank؛ أي «رُتبة الصفحات». يُعتقَد أن اسم الخوارزمية «بيدج رانك» قد اختير تيمُّنًا بلاري بيدج، أحد مؤسسي شركة جوجل، الذي نشر مقالات، بالتعاون مع الشريك المؤسِّس سيرجي برين، عن هذه الخوارزمية الجديدة. حتى صيف عام ٢٠١٦، كانت نتائج خوارزمية «بيدج رانك» متاحةً للجمهور عن طريق تنزيل شريط الأدوات «بيدج رانك». كانت أداة «بيدج رانك» العامة تعتمد على مقياس من ١ إلى ١٠. وقبل أن يتم حجبها، تَمكَّنتُ من حفظ بعض النتائج. إذا كتبتُ عبارة «البيانات الضخمة» في محرك بحث جوجل باستخدام الكمبيوتر المحمول، تصلني رسالة تخبرني بأنه يوجد «حوالي ٣٧٠ مليون نتيجة (في غضون ٠,٤٤ ثانية)» برتبة صفحات مقدارها ٩. وفي أعلى هذه القائمة، توجد بعض الإعلانات المدفوعة، تليها نتيجة البحث الخاصة بموقع ويكيبيديا. يترتَّب على البحث عن كلمة «بيانات» عرض ٥,٥٣٠,٠٠٠,٠٠٠ نتيجة في غضون ٠,٤٣ ثانية برتبة صفحات مقدارها ٩. ومن الأمثلة الأخرى، التي كانت جميعها برتبة صفحات مقدارها ١٠، موقع الحكومة الأمريكية، وفيسبوك، وتويتر، ورابطة الجامعات الأوروبية.

يعتمد أسلوب حساب رتبة الصفحات هذا على عدد الروابط المؤدية إلى صفحة ويب ما، فكلما زاد عدد الروابط، ارتفعت درجة التقييم، وظهرت الصفحة في مكان أكثر تقدمًا ضمن نتائج البحث. ولا يعكس هذا عدد مرات زيارة الصفحة. إذا كنت مصمِّم مواقع إلكترونية، فإنك ترغب في تحسين موقعك حتى يتصدَّر قائمة نتائج البحث بكلمات بحثٍ معينة؛ وذلك لأن أغلب الناس لا ينظرون إلى ما هو أبعد من نتائج البحث الثلاث أو الأربع الأولى. وهذا يتطلَّب عددًا هائلًا من الروابط، ويؤدي، لا محالة، إلى عملية متاجرة بالروابط. حاولت جوجل حل مشكلة الترتيب «الزائف» تلك عن طريق تعيين رتبة جديدة هي صفر للشركات المتورطة في الأمر، أو حتى إزالتها تمامًا من محرك بحث جوجل، إلا أن هذا لم يحلّ المشكلة، بل أجبر هذه التجارة على العمل في الخفاء، واستمر بيع الروابط.

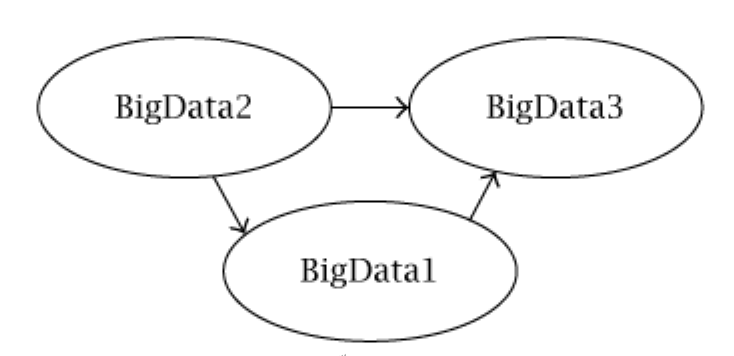
لم تُستبعَد خوارزمية «بيدج رانك» نفسها، بل أصبحت جزءًا من مجموعة كبيرة من برامج الترتيب غير المتاحة للعامة. يعيد محرك بحث جوجل حساب الرُّتَب بصورة دورية، بما يعكس الروابط المضافة وكذلك المواقع الإلكترونية الجديدة. وبما أن خوارزمية «بيدج رانك» حسَّاسة من الناحية التجارية، فلا توجد تفاصيل كاملة عنها متاحة للعامة، ولكن يمكننا تكوين فكرة عامة عنها بالنظر إلى مثال. تقدِّم الخوارزمية طريقةً معقدة لتحليل الروابط بين صفحات الويب بِناءً على نظرية

الاحتمالات، حيث تَشير الاحتمالية «واحد» إلى اليقين والاحتمالية «صفر» إلى الاستحالة، وكل شيء آخر يحمل قيمةً احتمالية تتراوح بين هاتَين القيمتَين.

لفهم كيفية تحديد الرُّتَب، نحتاج أولًا إلى أن نعرف الشكل الذي يكون عليه التوزيع الاحتمالي. إذا فكَّرنا في نتيجة إلقاء نرد ذي ستة أوجه متساوية، فإن النتائج من ١ إلى ٦ تحمل احتمالية الظهور نفسها، وعليه، فإن كلًّا منها له احتمالية بنسبة ٦/١. تصف القائمةُ التي تتضمَّن جميع النتائج المحتملة، بالإضافة إلى احتمالية حدوث كل منها، التوزيعَ الاحتمالي.

بالرجوع مرةً أخرى إلى مسألة ترتيب صفحات الويب حسب الأهمية، لا يمكننا القول إن جميعها متساوية من حيث الأهمية، ولكن إذا توافرت لنا طريقة لتعيين الاحتمالات لكل صفحة ويب، فمن شأن هذا أن يمنحنا مؤشرًا معقولًا عن مدى أهميتها. إذن، ما تفعله خوارزميات على غرار «بيدج رانك» هو أنها تنشئ توزيعًا احتماليًّا لشبكة الويب بأكملها. لتفسير ذلك، دعونا نتخيَّل متصفحًا عشوائيًّا للويب يبدأ التصفح من أي صفحة ويب ثم ينتقل إلى صفحة أخرى باستخدام الروابط المتاحة.

سنتناول مثالًا مُبسَّطًا يتضمَّن موقعًا إلكترونيًّا مكوَّنًا من ثلاث صفحات ويب فقط، وهي BigData1،وBigData2، وBigData3. لنفترض أن الروابط الوحيدة توجد ما بين BigData2 وBigData3، وما بين BigData2 وBigData1، وما بين BigData1 وBigData3. يمكن إذن تمثيل هذا الموقع الإلكتروني على النحو الموضَّح في شكل ٤-٣، حيث تُمثِّل العُقَد صفحات الويب وتُمثِّل الأسهم (الأضلاع) الروابط.

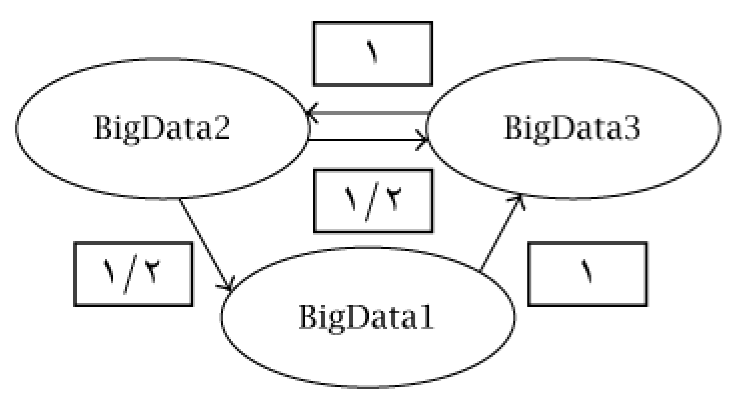


شكل ٤-٣: رسم بياني موجَّه يُمثِّل جزءًا صغيرًا من الموقع الإلكتروني.

لكل صفحة رُتبة تدل على مدى أهميتها أو شيوعها. ستكون صفحة BigData3 هي الأعلى رُتبة؛ لأن أغلب الروابط تتجه إليها، ما يجعلها الأكثر شيوعًا. والآن، لنفترض أن متصفحًا عشوائيًّا يزور صفحة ويب، ومتاح له تصويتٌ نسبي بواقع صوتٍ واحد فقط للإدلاء به، والذي يُقسَّم بالتساوي بين اختياراته التالية من صفحات الويب. على سبيل المثال، إذا كان المتصفح العشوائي يزور حاليًّا صفحة BigData1، فإن الخيار الوحيد المتوفر أمامه هو زيارة صفحة BigData3 بعد ذلك.

وعليه، يمكننا القول إنه أُجري تصويتٌ بواقع صوتٍ واحد من قِبل BigData11 لصالح BigData4.

تُنشَأ روابط في موقع الويب الحقيقي باستمرار؛ ومن ثَمَّ، لنفترض أننا وجدنا الآن أن صفحة BigData3 تشتمل على رابط يؤدي إلى صفحة BigData2، كما هو موضَّح في شكل ٤-٤، إذن ستتغيَّر قيمة PageRank لصفحة BigData2؛ لأن المتصفح العشوائي أصبح لديه الآن أكثر من وجهة للانتقال إليها بعد صفحة BigData3.



شكل ٤-٤: رسم بياني موجَّه يُمثِّل جزءًا صغيرًا من الموقع الإلكتروني مع إضافة الرابط.

إذا بدأ المتصفح العشوائي في مثالنا الحالي عند صفحة BigData1، وكان الخيار الوحيد أمامه هو الانتقال إلى صفحة BigData3 بعدها، فإن التصويت بأكمله بواقع صوت واحد ينتقل إلى BigData3، وتحصل صفحة BigData2 على صفر من الأصوات. أمَّا إذا بدأ المتصفح العشوائي عند صفحة BigData2، فسيُقسَّم التصويت بالتساوي بين الصفحتَين BigData3 وBigData1. وأخيرًا، إذا بدأ المتصفح العشوائي عند BigData3، فسوف يتنقَّل عدد الأصوات كاملًا إلى BigData2. يعرض جدول ٤-٤ مُلخَّصًا بقيم «التصويت» النسبي هذه.

باستخدام جدول ٤-٤، يمكننا أن نرى الآن إجمالي عدد الأصوات المُدلى بها لصالح كل صفحة ويب كالآتي:

إجمالي الأصوات لصالح BD1 هو ١/٢ (من قِبل BD2).

إجمالي الأصوات لصالح BD2 هو ١ (من قِبل BD3).

إجمالي الأصوات لصالح BD3 هو ١ ١/٢ (من قِبل BD1 وBD2).

جدول ٤-٤: الأصوات المُعطاة لكل صفحة ويب

	نسبة الأصوات المُعطاة من قِبل BD1	نسبة الأصوات المُعطاة من قِبل BD2	نسبة الأصوات المُعطاة من قِبل BD3
لصالح BD1	صفر	٢/١	صفر
لصالح BD2	صفر	صفر	١
لصالح BD3	١	٢/١	صفر

وبما أن اختيار صفحة البدء يكون عشوائيًّا، فإن احتمالية اختيار المتصفح لكل صفحة منها يكون متساويًا؛ ومن ثمَّ تُعيَّن لكل منها رُتبة صفحة مبدئية هي ٣/١. لتحديد رتب الصفحات المرغوب فيها فيما يخص مثالنا الحالي، علينا أن نُحدِّث رُتب الصفحات المبدئية حسب نسبة الأصوات المُعطاة لكل صفحة.

على سبيل المثال، حصلت صفحة BD1 على ٢/١ صوت، أعطته لها صفحة BD2، وعليه، فإن رُتبة صفحة BD1 هي ٣/١ × ٢/١ = ٦/١. وبالمثل، تُحسَب رُتبة صفحة BD2 من خلال ٣/١ × ١ = ٦/٢، ورُتبة BD3 من خلال ٣/١ × ٢/٣ = ٦/٣. وبما أن رُتب الصفحات مجموعها يساوي واحدًا، نلجأ إلى التوزيع الاحتمالي الذي يُحدِّد أهمية كل صفحة أو رُتبتها.

ولكننا سنواجه عقبةً هنا. قلنا سابقًا إنَّ احتمالية أن يبدأ متصفحٌ عشوائي التصفح من أي صفحة تساوي ٣/١. وبعد خطوة واحدة، حسبنا أن احتمالية بدء المتصفح العشوائي التصفح من صفحة BD1 تساوي ٦/١. ماذا سيحدث بعد خطوتَين؟ حسنًا، مرةً أخرى نستخدم رُتب الصفحات الحالية كأصواتٍ لحساب رُتب الصفحات الجديدة. ستكون العمليات الحسابية مختلفةً قليلًا في هذه المرحلة؛ لأن رُتب الصفحات الحالية ليست متساوية، ولكن الطريقة لم تتغيَّر، ما يعطينا رُتب صفحات جديدة كالآتي: رُتبة صفحة BD1 هي ١٢/٢، ورُتبة صفحة BD2 هي ١٢/٦، ورُتبة صفحة BD3 هي ١٢/٤. تُكرَّر هذه الخطوات، أو التكرارات، حتى تتقارب الخوارزمية، وهذا يعني أن العملية تستمر على هذا المنوال حتى لا يكون هناك مجال لإجراء أي تغييرات أخرى بِناءً على أي عمليات ضرب أخرى. وبعد الوصول إلى الترتيب النهائي، يمكن لخوارزمية «بيدج رانك» أن تختار الصفحة ذات الرتبة الأعلى لعملية بحث معينة.

قدَّم بيدج وبرين، في أوراقهما البحثية الأصلية، معادلةً لحساب رُتب الصفحات، تضمَّنت معامل تخميد d، والذي يُعرَّف بأنه احتمالية أن ينقر متصفح ويب عشوائي على أحد الروابط في الصفحة الحالية. ومن ثمَّ، فإن احتمالية عدم نقر متصفح ويب عشوائي على أحد الروابط في الصفحة الحالية تساوي $(1 - d)$، ما يعني أن المتصفح العشوائي قد أنهى التصفح. ضمِن معامل التخميد هنا أن ينتهي الحال بمتوسط رُتب الصفحات على مستوى الموقع الإلكتروني بالكامل عند ١، بعد إجراء عددٍ كافٍ من الحسابات التكرارية. قال بيدج وبرين إن متوسط رتب الصفحات في موقع إلكتروني مُكوَّن من ٣٢٢ مليون رابط تحدَّد بعد ٥٢ تكرارًا.

مجموعات البيانات العامة

ثمَّة الكثير من مجموعات البيانات الضخمة المتاحة مجانًا، والتي يمكن أن يستخدمها الأفراد المهتمون أو المجموعات المهتمة في مشروعاتهم. وتُعَد مؤسسة كومون كراول، التي ذكرناها في موضع سابق في هذا الفصل، مثالًا على ذلك. تضمَّن الأرشيف الشهري لمؤسسة كومون كراول، الذي يستضيفه برنامج أمازون لمجموعات البيانات العامة، في أكتوبر ٢٠١٦، ما يزيد على ٣,٢٥ مليار صفحة ويب. تتضمَّن مجموعات البيانات العامة مجموعةً كبيرة من التخصُّصات، بما في ذلك بيانات الجينوم، وصور الأقمار الصناعية، وبيانات الأخبار العالمية. وبالنسبة إلى أولئك الذين من غير المرجَّح أن يكتبوا النصوص البرمجية بأنفسهم، توفِّر أداة جوجل للتحليل الإحصائي للكلمات (Google Ngram Viewer) طريقةً مشوِّقة لاستكشاف عددٍ من مجموعات البيانات الضخمة على نحو تفاعلي (انظر جزء «قراءات إضافية» لمعرفة التفاصيل).

نموذج البيانات الضخمة

رأينا سابقًا بعضًا من طرق الاستفادة من البيانات الضخمة، وتحدَّثنا في الفصل الثاني عن البيانات الصغيرة. بالنسبة إلى تحليل البيانات الصغيرة، يمكن استخدام الأسلوب العلمي على نحو راسخ تمامًا وينطوي بالضرورة على التفاعل البشري: شخصٌ تتراءى لذهنه فكرة ما، ثم يضع فرضيةً أو نموذجًا فكريًّا، ويبتكر طرقًا لاختبار توقعاته. كتبَ عالِمُ الإحصاء الشهير جورج بوكس عام ١٩٧٨: «جميع النماذج خاطئة، ولكن بعضها مفيد». وما يعنيه بهذه العبارة أنَّ النماذج الإحصائية والعلمية، بوجه عام، لا تُقدِّم تمثيلاتٍ دقيقةً للعالَم من حولنا، ولكن يمكن لنموذج فكري جيد أن يُقدِّم تصوُّرًا مفيدًا لِمَا يجب أن تستند إليه التوقعات ويستخرج النتائج بطريقة موثوقة. ولكن، كما أوضحنا سابقًا، فإننا لا نتَّبع هذه الطريقة عند التعامل مع البيانات الضخمة. بدلًا من ذلك، نجد أن السيادة للآلة وليس للعالِم.

وصفَ توماس كون، في إحدى كتاباته عام ١٩٦٢، مفهوم الثورات العلمية التي تلي فتراتٍ طويلةً من العلم العادي عندما يُطوَّر نموذج حالي ويُدرَس من جميع جوانبه. وإذا ظهر عدد كافٍ من الانحرافات التي لا يمكن حلها وتؤدي إلى تقويض أركان نظريةٍ قائمة، ما يؤدي بالباحثين إلى فقدان الثقة فيها، فإن هذا يُسمَّى «أزمة»، وتُحَل في نهاية المطاف بوضع نظرية جديدة أو نموذج فكري جديد. ولكي يُقبَل نموذج فكري جديد، فإنه لا بد أن يُجيب عن بعض الأسئلة الإشكالية الموجودة في النموذج الفكري القديم. ولكن، بوجه عام، لا يطمس النموذج الجديد النموذج السابق بالكامل. على سبيلِ المثال، غيَّر التحوُّل من ميكانيكا نيوتن إلى النظرية النسبية لأينشتاين من نظرة العِلم إلى العالَم، دون أن يطرح قوانين نيوتن جانبًا: تُمثِّل حاليًّا ميكانيكا نيوتن حالةً خاصة من نظرية النسبية الأوسع نطاقًا. كما يُمثِّل التحوُّل من علم الإحصاء الكلاسيكي إلى أساليب تحليل البيانات الضخمة تغيُّرًا كبيرًا، وتجتمع فيه الكثير من السمات المميِّزة للتحوُّل النوعي. وعليه، فإنَّ الأمر يستلزم حتمًا تطوير أساليب للتعامل مع هذا الوضع الجديد.

دعونا نتناول أسلوب إيجاد ارتباطات في البيانات الضخمة، والذي يوفِّر وسيلةً للتوقع بناءً على قوة العلاقات بين المتغيِّرات. من المتعارَف عليه في علم الإحصاء الكلاسيكي أنَّ الارتباط لا يقتضي السببية. على سبيل المثال، قد يُسجِّل مُعلِّم عدد مرات غياب أحد الطلاب عن المحاضرات ودرجات الطالب؛ ومن ثَمَّ، عندما يجد ارتباطًا واضحًا بينها، قد يستخدم غياب الطالب في توقع درجاته. ولكن، لن يكون من الصائب أن يستنتج أن عدد مرات غياب الطالب سبب في تدني درجاته. لا يمكننا معرفة السبب في ارتباط متغيِّرَين من خلال النظر إلى العمليات الحسابية المجرَّدة فحسب؛ فربما الطلاب الأقل قدرةً على الاستيعاب يميلون إلى التغيُّب عن الصف، وربما لا يمكن للطلاب الذين يغيبون بسبب المرض أن يعوِّضوا ما فاتهم لاحقًا. ومن ثَمَّ، لا بد من التفاعل والتفسير البشري لتحديد أي الارتباطات مفيدة.

فيما يخصُّ البيانات الضخمة، يؤدي استخدام الارتباطات إلى ظهور مشكلاتٍ إضافية. فإذا تناولنا مجموعة بيانات هائلة، يمكن كتابة خوارزميات تؤدي — عند تطبيقها — إلى عدد كبير من الارتباطات الزائفة، التي تكون مستقلةً تمامًا عن وجهات نظر أي إنسان أو آرائه أو فرضياته. تنشأ مشكلات بسبب الارتباطات الزائفة — على سبيل المثال، معدلات الطلاق واستهلاك السمن النباتي، وهي أحد الارتباطات الزائفة الكثيرة التي تحدَّثت عنها وسائل الإعلام. يمكننا أن نرى مدى سُخف هذا الارتباط من خلال تطبيق الأسلوب العلمي. ولكن، عندما يصبح عدد المتغيِّرات كبيرًا، يزداد أيضًا عدد الارتباطات الزائفة. تُعد هذه إحدى المشكلات الرئيسية المصاحبة لمحاولة استخراج معلومات مفيدة من البيانات الضخمة؛ لأننا عندما نفعل ذلك، مثلما هو الحال مع التنقيب في البيانات الضخمة، فإننا عادةً ما نبحث عن أنماط وارتباطات. وكما سنرى في الفصل الخامس، كانت هذه المشكلات هي أحد أسباب فشل توقعات خدمة «اتجاهات الأنفلونزا من جوجل».

الفصل الخامس

البيانات الضخمة والطب

غيَّر تحليل البيانات الضخمة من مجال الرعاية الصحية إلى حَدٍّ كبير. لم تُدرَك كامل إمكانات هذا المجال بعد، ولكنه يشمل التشخيص الطبي، وبالتنبؤ بالأوبئة، وقياس الاستجابة العامة للتحذيرات الصحية الحكومية، وتقليل التكاليف المرتبطة بأنظمة الرعاية الصحية. ولنبدأ بتناول ما أصبح يُسمَّى اصطلاحًا «معلوماتية الرعاية الصحية».

معلوماتية الرعاية الصحية

تُستخدَم الأساليب العامة التي تحدَّثنا عنها في الفصول السابقة في جمع البيانات الطبية الضخمة، وتخزينها، وتحليلها. وبوجه عام، تَستخدم معلوماتية الرعاية الصحية وفروعها المعرفية العديدة، مثل المعلوماتية السريرية والمعلوماتية الحيوية، البياناتِ الضخمة لتقديم رعاية مُحسَّنة للمرضى وتقليل التكاليف. لنتأمَّل معايير تعريف البيانات الضخمة (التي ناقشناها في الفصل الثاني) — الحجم، والتنوُّع، والسرعة، والموثوقية — ونرى كيف تنطبق على البيانات الطبية. يتحقَّق معيارَا الحجم والسرعة، على سبيل المثال، عند جمع البيانات المتعلقة بالصحة العامة عبر مواقع شبكات التواصل الاجتماعي من أجل تتبُّع مسار الأوبئة، ويتحقَّق معيار التنوع عند تخزين سجلات المرضى بالتنسيق النصي، سواءٌ أكان هيكليًّا أم غير هيكلي، وكذلك عند جمع بيانات أجهزة الاستشعار مثل البيانات التي تُوفِّرها أجهزة التصوير بالرنين المغناطيسي، ويُعَد معيار الموثوقية معيارًا أساسيًّا في الاستخدامات الطبية، ويُولي عنايةً فائقة بإزالة البيانات غير الدقيقة.

قد تكون وسائل التواصل الاجتماعي مصدرًا قيمًا للمعلومات ذات الصلة في المجال الطبي، وذلك من خلال جمع البيانات من مواقع مثل فيسبوك وتويتر والمدوَّنات المختلفة، ولوحات تبادل الرسائل، وعمليات البحث على الإنترنت. يوجد الكثير من لوحات تبادل الرسائل التي تركِّز على موضوعات معينة في مجال الرعاية الصحية، حيث تُقدِّم ثروةً من البيانات غير الهيكلية. جُمِعَت منشورات على كلٍّ من موقعَي فيسبوك وتويتر، باستخدام أساليب تصنيف مماثلة لتلك التي شرحناها في الفصل الرابع، لرصد تجربة التفاعلات غير المرغوبة تجاه الأدوية وتزويد مختصي الرعاية الصحية بمعلوماتٍ مفيدة بشأن تفاعلات الأدوية وإساءة استخدامها. أصبح التنقيب في البيانات عبر وسائل التواصل الاجتماعي، لأغراض البحث في مجال الصحة العامة، ممارسةً مُعترَفًا بها في المجتمع الأكاديمي.

تُقدِّم مواقع شبكات التواصل الاجتماعي المُخصَّصة للعاملين في المجال الطبي، مثل سيرمو إنتيليجنس، وهي شبكة طبية عالمية تصف نفسها بأنها «أكبر شركة عالمية في مجال جمع بيانات الرعاية الصحية»، لمختصي الرعاية الصحية فوائدَ فوريةً من حشد المصادر يكتسبونها من التعامل

مع أقرانهم. تتزايد شهرة مواقع تقديم المشورة الطبية عبر الإنترنت؛ ومن ثمَّ فإنها تنشئ المزيد من المعلومات. ربما تكون مجموعة «السجلات الصحية الإلكترونية» المصدر الأهم بين تلك المواقع، وإن كانت إمكانية الوصول إليها غير متاحة للعامة. توفِّر هذه السجلات نسخةً إلكترونية من التاريخ الطبي الكامل للمريض، بما في ذلك التشخيصات الطبية، والأدوية الموصوفة، وصور الأشعة الطبية مثل أشعة إكس، وجميع المعلومات الأخرى ذات الصلة التي جُمعَت بمرور الزمن، ومن ثم إنشاء «مريض افتراضي»، وهو مفهوم سنتناوله لاحقًا في هذا الفصل. بالإضافة إلى استخدام البيانات الضخمة في تحسين رعاية المرضى وتقليل التكاليف، من خلال جمع المعلومات المتولِّدة من مجموعة متنوِّعة من المصادر عبر الإنترنت، أصبح من الممكن التفكير في التنبؤ بمسار الأوبئة الحديثة الظهور.

خدمة اتجاهات الأنفلونزا من جوجل

تواجه الولايات المتحدة كلَّ عام، شأن الكثير من الدول، وباء الأنفلونزا الذي يؤدي إلى زيادة الأعباء على الموارد الطبية وارتفاع الخسائر في الأرواح. تُمثِّل بياناتُ الأوبئة السابقة المقدَّمة من المركز الأمريكي لمكافحة الأمراض، وهو الوكالة المنوطة بمراقبة الصحة العامة، بالإضافة إلى أساليب تحليل البيانات الضخمة، القوةَ الدافعة لجهود الباحثين الرامية إلى التنبؤ بانتشار الأمراض من أجل تركيز الخدمات والحَدِّ من انتشار هذه الأمراض.

بدأ فريقُ خدمة اتجاهات الأنفلونزا العملَ على التنبؤ بأوبئة الأنفلونزا باستخدام بيانات محركات البحث. وانصبَّ عمله على الكيفية التي يمكن بها التنبؤ بوباء الأنفلونزا السنوي في فترة زمنية أقصر ممَّا يستغرقها حاليًّا مركز مكافحة الأمراض في معالجة بياناته. في رسالة نُشرَت في مجلة «نيتشر» العلمية المرموقة في فبراير ٢٠٠٩، شرحَ الفريقُ المُكوَّن من ستة من مهندسي البرامج لدى شركة جوجل ما كانوا بصدد تنفيذه. إذا أمكن استخدام البيانات في التنبؤ على نحو دقيق بمسار وباء الأنفلونزا السنوي في الولايات المتحدة، لأمكنَ احتواءُ المرض، وإنقاذ الأرواح، وتوفير الموارد الطبية. تطرَّقَ فريقُ جوجل إلى فكرة أن هذا الأمر يمكن تحقيقه من خلال جمع استعلامات محركات البحث بشأن المخاوف المتعلقة بمرض الأنفلونزا وتحليلها. آلت المحاولات السابقة لاستخدام بيانات الإنترنت في التنبؤ بانتشار الأنفلونزا إلى أحدِ أمرَين، لا ثالثَ لهما: إمَّا أنها باءت بالفشل، وإمَّا أنها حقَّقت نجاحًا محدودًا. ولكن، من خلال التعلُّم من الأخطاء السابقة في هذا البحث الرائد، راودَ الأملُ شركة جوجل ومركز مكافحة الأمراض في نجاح تجربة استخدام البيانات الضخمة المتولِّدة من استعلامات محركات البحث في تتبع مسار الوباء.

يجمع مركز مكافحة الأمراض، ونظيره البرنامج الأوروبي لرصد الأنفلونزا، البيانات من مختلِف المصادر، بما في ذلك الأطباء الذين يقدِّمون تقارير بأعداد المرضى الذين يعالجونهم ولديهم أعراض شبيهة بالأنفلونزا. ولكن، بحلول الوقت الذي يتم فيه دمج هذه البيانات، يكون قد مرَّ عليها عادةً أسبوعان، ويكون الوباء قد ازداد انتشارًا. باستخدام البيانات المُجمَّعة في الوقت الحقيقي من الإنترنت، كان فريقُ شركة جوجل ومركز مكافحة الأمراض يهدفان إلى تحسين دقة التنبؤات

الخاصة بالوباء والتوصل إلى نتائج في غضون يوم واحد. ولتحقيق ذلك، جُمعَت بيانات حول استعلامات البحث المتعلِّقة بالأنفلونزا، والتي تراوحت ما بين البحث الفردي على الإنترنت عن طرق علاج الأنفلونزا وأعراضها، والبيانات الوفيرة مثل الاتصالات الهاتفية التي أُجريت بمراكز تقديم المشورة الطبية. واستطاعت جوجل الوصول إلى كمية هائلة من بيانات استعلامات البحث التي تجمَّعت لديها خلال الفترة ما بين عامَي ٢٠٠٣ و٢٠٠٨، ومن خلال استخدام عناوين «آي بي»، أمكنَ تحديد الموقع الجغرافي الذي نشأت منه استعلامات البحث؛ ومن ثَمَّ، تصنيف البيانات في مجموعاتٍ حسب الولاية. جُمِعَت بيانات مركز مكافحة الأمراض من عشر مناطق، تتضمَّن كلٌّ منها البيانات التراكمية من مجموعة من الولايات (على سبيل المثال، تشمل المنطقة التاسعة ولايات أريزونا، وكاليفورنيا، وهاواي، ونيفادا)، وتُدمَج هذه البيانات بعد ذلك في النموذج.

اعتمدَ مشروع اتجاهات الأنفلونزا من جوجل على النتيجة المعروفة بأن ثمَّة ارتباطًا وثيقًا بين عدد عمليات البحث المتعلقة بالأنفلونزا على شبكة الإنترنت وعدد زيارات عيادات الأطباء. فإذا كان ثمة عدد كبير من الأشخاص في منطقة معينة يبحثون عن معلوماتٍ متعلِّقة بالأنفلونزا على شبكة الإنترنت، فربما أصبح بالإمكان توقُّع انتشار حالات الإصابة بالأنفلونزا في المناطق المجاورة لها. وبما أن الاهتمام ينصَب على تقدير الاتجاهات، أصبح من الممكن تجهيل البيانات؛ ومن ثَمَّ انتفت ضرورة الحصول على موافقة الأفراد. وباستخدام بيانتها التراكمية على مدار خمس سنوات، والتي اقتصرت على الإطار الزمني نفسه لبيانات مركز مكافحة الأمراض؛ ومن ثَمَّ جُمعَت خلال موسم الأنفلونزا فقط، حسبت جوجل معدل التكرار الأسبوعي لكل استعلام من استعلامات البحث الأكثر شيوعًا، البالغ عددُها ٥٠ مليونًا، التي تغطِّي جميع الموضوعات. ثم قُورنت أعداد استعلامات البحث هذه ببيانات مركز مكافحة الأمراض الخاصة بالأنفلونزا، واستُخدِمَت البيانات ذات الارتباط الأعلى في نموذج تقدير اتجاهات الأنفلونزا. اختارت جوجل استخدام أول ٤٥ مصطلحًا من مصطلحات البحث المتعلقة بالأنفلونزا وأكثرها تكرارًا، ثم تتبَّعتها في استعلامات البحث التي يُجريها الناس. وعلى الرغم من سرية القائمة الكاملة لمصطلحات البحث، فإنها تشمل، على سبيل المثال لا الحصر، «مضاعفات الأنفلونزا»، و«علاج نزلة البرد/الأنفلونزا»، و«الأعراض العامة للأنفلونزا». شكَّلت البيانات التاريخية خطًّا مرجعيًّا يُقيَّم على أساسه تأثير الأنفلونزا الحالي على مصطلحات البحث المختارة، وبمقارنة بيانات الوقت الحقيقي الجديدة بهذه البيانات، وُضِع على مقياس من ١ إلى ٥، حيث يعني العدد ٥ الأكثر خطورة.

استُخدِمَت خوارزمية «جوجل للبيانات الضخمة» في موسمَي الأنفلونزا للعام ٢٠١١-٢٠١٢ والعام ٢٠١٢-٢٠١٣، ولكنها فشلت في تحقيق أهدافها. وبعد انتهاء موسم الأنفلونزا، قُورنت تنبؤاتها بالبيانات الفعلية لمركز مكافحة الأمراض. خلال إنشاء النموذج، الذي كان يجدر به أن يكون تمثيلًا جيدًا لاتجاهات الأنفلونزا المستقاة من البيانات المتوافرة، جاءَ عددُ حالات الإصابة بالأنفلونزا الذي توصَّلت إليه خوارزمية «اتجاهات الأنفلونزا من جوجل» مبالغًا فيه؛ حيث فاقَ العددَ الفعلي بنسبة ٥٠ بالمائة على الأقل خلال الأعوام التي استُخدمَت فيها الخوارزمية. توجد أسباب عدة لعدم تحقيق هذا النموذج النجاح المأمول. فقد استُبعِدَت بعضُ مصطلحات البحث عَمدًا؛ لأنها لم توافق توقُّعات فريق البحث. والمثال الأشهر والأكثر تداولًا على ذلك هو أن رياضة كرة السلة في المدارس الثانوية، التي يبدو أنها لا علاقة لها بالأنفلونزا، كانت رغم ذلك مرتبطةً ارتباطًا وثيقًا ببيانات مركز مكافحة الأمراض، ولكنها استُبعِدَت من النموذج. دائمًا ما تُمثِّل عملية اختيار المتغيِّر، وهي العملية

التي تَختار فيها عوامل التنبؤ الأكثر ملاءمة، مشكلةً مستعصية؛ ومن ثمَّ فإنها تَجرَى باستخدامِ الخوارزميات تجنُّبًا للانحياز. حافظت جوجل على سرية التفاصيل الخاصة بخوارزميتها، مشيرةً فقط إلى أنَّ رياضة كرة السلة في المدارس الثانوية قد حلَّت ضمن أعلى ١٠٠ مصطلح بحث استخدامًا، وبرَّرت استبعادها بتوضيح أن كلًّا من الأنفلونزا وكرة السلة يبلغان الحد الأقصى لمعدَّل الاستخدام في الوقت نفسه من العام.

كما أشرنا سابقًا، استخدمت جوجل خلال إنشاء النموذج الخاص بها ٤٥ مصطلح بحث لتكون بمثابة عوامل تنبؤ بالأنفلونزا. ولو أنها استخدمت مصطلحَ بحثٍ واحدًا، كـ «الأنفلونزا» على سبيل المثال، لكانت معلوماتٍ مهمةً وذات صلة، مثل جميع عمليات البحث المتعلِّقة بـ «علاج نزلات البرد»، قد مرَّت دون ملاحظة أو توثيق. تزداد دقة التنبؤ مع الاستعانة بعددٍ كافٍ من مصطلحات البحث، ولكن من الوارد أن تقل أيضًا إذا كان عدد مصطلحات البحث أكثر من اللازم. تُستخدَم البيانات الحالية كبيانات تدريبية لإنشاء نموذج يمكنه التنبؤ باتجاهات البيانات المستقبلية، ونظرًا لوجود عدد كبير للغاية من عوامل التنبؤ، لا يُضمَّن في البيانات التدريبية للنموذج سوى حالاتٍ عشوائية قليلة؛ ومن ثمَّ، فإنه على الرغم من أن النموذج يتوافق جيدًا مع البيانات التدريبية، فإنه لا يمكنه تقديم تنبؤاتٍ جيدة. يبدو أن هذه الظاهرة المتناقضة، التي تُسمَّى «الملاءمة المفرطة»، لم يضعها فريق العمل في اعتبارهم بما يكفي. ربما كان استبعادُ كرة السلة في المدارس الثانوية كأحد عوامل التنبؤ بسبب أنه يتزامن مع موسم الأنفلونزا أمرًا منطقيًّا، إلا أن ثمَّة ٥٠ مليون مصطلح بحثٍ آخر، ومع وجود هذا العدد الكبير فإنه لأمرٌ شبه محتَّم أن ترتبط مصطلحاتُ بحثٍ أخرى ارتباطًا وثيقًا ببيانات مركز مكافحة الأمراض، ولكنها لن تكون ذات صلة باتجاهات الأنفلونزا.

يتردَّد على عيادات الأطباء أشخاصٌ يعانون أعراضًا شبيهة بالأنفلونزا، إلا أن التشخيص غالبًا لا يكون الإصابة بالأنفلونزا (ربما يكون — مثلًا — نزلة برد عادية). أفرزت البيانات التي استخدمتها جوجل، وجمعتها على نحو انتقائي من استعلامات محرِّك البحث، نتائج غير سليمة من الناحية العلمية جراء التحيُّز الواضح، الذي نتجَ — على سبيل المثال — من استبعاد كلِّ مَن لا يستخدمون أجهزة الكمبيوتر وكلِّ مَن يستخدمون محركات بحثٍ أخرى. وثمَّة مشكلة أخرى ربما ساهمت في هذه النتائج غير الدقيقة، وهي أن العملاء الذين يبحثون في محرك بحث جوجل عن «أعراض الأنفلونزا» ربما تصفَّحوا بالفعل عددًا من المواقع الإلكترونية المتعلقة بالأنفلونزا، ممَّا أدَّى إلى حساب مرات استخدام هذا المصطلح وحده من مصطلحات البحث أكثر من مرة؛ ومن ثمَّ أدَّى إلى تفاقُم الأعداد. علاوةً على ذلك، يتغيَّر سلوك البحث بمرور الوقت، لا سيَّما في فترات تفشي الأوبئة، ولا بد من وضع هذا الأمر في الاعتبار عن طريق تحديث النموذج بصفة دورية. عندما تبدأ أخطاء التنبؤ في الظهور، فإنها تميل إلى التتابع، وهذا ما حدث مع تنبؤات «اتجاهات الأنفلونزا من جوجل»: انتقلت أخطاءُ أسبوعٍ ما إلى الأسبوع الذي يليه. دُرِسَت استعلامات البحث كما ظهرت بالفعل، ولم تُصنَّف في مجموعاتٍ حسب الهجاء أو الصياغة. وكان المثال الذي قدَّمته جوجل على ذلك هو أن كلًّا من عباراتٍ «دلائل الأنفلونزا»، و«الدلائل على الأنفلونزا»، و«الدلائل على مرض الأنفلونزا» قد أُحصي كلٌّ منها على حدة.

تعرَّض البحث، الذي يرجع تاريخه إلى موسم ٢٠٠٧-٢٠٠٨، للكثير من الانتقادات، التي كان بعضُها متحامِلًا، إلا أن الانتقادات كانت تتعلَّق عادةً بافتقار الشفافية، على سبيل المثال، رفضُ

الكشف عن كلِّ مصطلحات البحث المختارة والإحجام عن قَبول الطلبات المقدَّمة من المجتمع الأكاديمي للحصول على معلومات. إن بيانات استعلامات محرك البحث ليست نتاج تجربة إحصائية مخطَّطة، كما أنَّ إيجاد طريقة لتحليل هذه البيانات على نحو مُجدٍ واستخراج معلوماتٍ مفيدة منها يُعد مجالًا جديدًا ومليئًا بالتحديات قد يستفيد من التعاون. في موسم ٢٠١٢-٢٠١٣، أدخلت جوجل تغييراتٍ كبيرةً على خوارزمياتها، وبدأت في استخدام أسلوب رياضي جديد نسبيًّا يُسمَّى «الإلاستيكنت»؛ أي الشبكة المرنة، والذي يوفِّر وسيلةً دقيقة لاختيار عوامل التنبؤ اللازمة وتقليل عددها. عام ٢٠١١، بدأت جوجل مشروعًا مماثلًا لتتبُّع مسار حُمَّى الضَّنك، ولكنها لم تَعُد تنشر تنبؤاتٍ حيالها، وعام ٢٠١٥ تمَّ إيقاف مشروع اتجاهات الأنفلونزا من جوجل. ولكنها أصبحت الآن تشارك بياناتها مع الباحثين الأكاديميين.

قدَّم مشروع اتجاهات الأنفلونزا من جوجل، إحدى المحاولات الأولى لاستخدام البيانات الضخمة في التنبؤ بالأوبئة، أفكارًا مفيدة للباحثين الذين شرعوا في عملهم بعد هذا المشروع. وعلى الرغم من أن نتائج المشروع لم ترقَ لمستوى التوقعات، فمن الوارد فيما يبدو أن تظهر في المستقبل طرقٌ أفضل، وعندئذٍ ستتحقَّق الإمكاناتُ الكاملة للبيانات الضخمة في مجال تتبع مسار الأوبئة. أُجريت إحدى هذه المحاولات على يد فريق من العلماء من مختبرٍ لوس ألاموس الوطني في الولايات المتحدة، باستخدام بياناتٍ من موسوعة ويكيبيديا. وفازَ فريقُ دلفي البحثي في جامعة كارنيجي ميلون بتحدي مركز مكافحة الأمراض تحت عنوان «تنبَّأ بالأنفلونزا» عن موسمَي ٢٠١٤-٢٠١٥ و٢٠١٥-٢٠١٦ لاختيار أفضل خبراء التنبؤ. نجحَ الفريقُ في استخدام بياناتٍ من جوجل وتويتر وويكيبيديا لمراقبة حالات تفشي الأنفلونزا.

تفشي وباء الإيبولا في غرب أفريقيا

شهدَ العالمُ قديمًا الكثيرَ من الأوبئة؛ فقد قتلت الأنفلونزا الإسبانية عامَي ١٩١٨-١٩١٩ ما يتراوح بين ٢٠ و٥٠ مليون نسمة، وبلغَ إجمالي عدد الإصابات وقتها نحو ٥٠٠ مليون نسمة. كانت المعلومات المتوافرة عن الفيروس قليلةً للغاية، ولم يكن هناك علاج مجدٍ، وكانت استجابة الصحة العامة محدودة، وهو أمرٌ يرجع بلا شك إلى نقص المعرفة. تغيَّر هذا الوضع عام ١٩٤٨ بالافتتاح الرسمي لمنظمة الصحة العالمية، التي تولَّت مسئولية مراقبة الصحة العالمية وتحسينها من خلال التعاون والتضافر بين دول العالم. في الثامن من أغسطس عام ٢٠١٤، في اجتماعٍ هاتفي عن بُعد للجنة الطوارئ المَعنيَّة باللوائح الصحية الدولية، أعلنت منظمة الصحة العالمية أن تَفَشي الإيبولا في غرب أفريقيا أصبح يشكِّل رسميًّا «طارئةً صحية عامة تثير قلقًا دوليًّا». وطبقًا للتعريف الذي قدَّمته منظمة الصحة العالمية لهذه العبارة، فإنَّ تفشي الإيبولا قد شكَّل «حدثًا استثنائيًّا» يستوجب جهودًا دولية غير مسبوقة لاحتوائه؛ ومن ثمَّ، تفادي حدوث وباء.

طَرَحَ تفشي الإيبولا في غربٍ أفريقيا في ٢٠١٤، الذي اقتصر في الأساس على دول غينيا وسيراليون وليبيريا، مجموعةً مختلفة من المشكلات مقارنةً بمشكلات تفشي وباء الأنفلونزا السنوي في الولايات المتحدة. كانت البيانات التاريخية عن فيروس الإيبولا إمَّا غير موجودة وإمَّا غير مفيدة؛

لأنه لم يُسجَّل من قبل تفشٍ بهذا الحجم لهذا الفيروس، وعليه، ظهرت الحاجة لوضع استراتيجيات جديدة للتعامل معه. وعلى ضوء معرفة تحركات السكان التي من شأنها أن تساعد العاملين في مجال الصحة العامة في مراقبة انتشار الأوبئة، كان يُعتقَد أنه يمكن استخدام المعلومات التي تمتلكها شركات الهواتف المحمولة في متابعة حركات السفر في المناطق الموبوءة، وتطبيق إجراءات، على غرار فرض قيود على السفر، من شأنها أن تحتوي الفيروس؛ ومن ثمَّ، إنقاذ الأرواح. كان من المفترض في نموذج التفشي في الوقت الحقيقي الناتج أن يتنبأ بالأماكن التي على الأرجح أن يتفشَّى فيها المرضُ بعد ذلك، ومن ثمَّ تركيز الموارد طبقًا لذلك.

إنَّ المعلومات الرقمية التي يمكن جمعها من الهواتف المحمولة أوليةٌ بعضَ الشيء؛ رقم هاتف كلٍّ من المتصل والمتصَل به، وموقع تقريبي للمتصل؛ فالاتصالات التي تُجرَى باستخدام الهاتف المحمول تُنشئ سِجلًا يمكن استخدامه في تقدير موقع المتصل بِناءً على برج الاتصالات المُستخدَم لكل اتصال. فرضَ الوصولُ إلى هذه البيانات عددًا من المشكلات: شكَّلت مسائل الخصوصية هاجسًا حقيقيًّا؛ نظرًا لإمكانية الاستدلال على الأشخاص الذين لم يوافقوا على تتبُّع مسار مكالماتهم وتحديد هُويتهم.

في بلدان غرب أفريقيا التي مُنيَت بتفشي الإيبولا، لم تكن كثافة استخدام الهواتف المحمولة متماثلة، حيث سُجِّلَت أقلُ النسب في المناطق الريفية الفقيرة. على سبيل المثال، كان ما يزيدُ قليلًا عن نصف العائلات في ليبيريا وسيراليون عام ٢٠١٣، وهما دولتان من الدول التي تأثرت تأثيرًا مباشرًا بتفشي الإيبولا عام ٢٠١٤، لديه هواتف محمولة، ومع ذلك كانت البيانات التي قدَّموها كافيةً لتتبع حركة سكان الدولتَين على نحو مفيد.

أُعطيت بعضُ البيانات التاريخية التي جرى جمعها من الهواتف المحمولة إلى مؤسسة فلومايندر، وهي مؤسسة غير ربحية مقرُّها السويد، تُكرِّس نشاطها للتعامل مع البيانات الضخمة بشأن مشاكل الصحة العامة التي تؤثِّر في دول العالم الأكثر فقرًا. عام ٢٠٠٨، كانت مؤسسة فلومايندر أولَ جهة تستخدم بيانات شركات اتصالات الهواتف المحمولة في تتبُّع حركة السكان في بيئة حافلة بالتحديات الطبية، وذلك ضمن مبادرة أطلقتها منظمة الصحة العالمية للقضاء على مرض الملاريا؛ ومن ثمَّ، كانت أحد الاختيارات البديهية للتعامل مع أزمة الإيبولا. استخدم فريقٌ دوليٌّ بارز البيانات التاريخية المُجَهلة في وضعِ خرائط لحركة السكان في المناطق الموبوءة بالإيبولا. لم تكن هذه البيانات التاريخية مُستخدَمةً على نطاق واسع؛ نظرًا لتغيُّر سلوكيات السكان في فترات الأوبئة، إلا أنها أعطت مؤشراتٍ قويةً عن الأماكن التي سيميل الناس إلى السفر إليها في حالات الطوارئ. وتُقدِّم سجلات نشاط أبراج الهواتف المحمولة تفاصيل عن أنشطة السكان في الوقت الحقيقي.

ومع ذلك، جاءت أرقام تنبؤات تفشي الإيبولا التي نشرتها منظمة الصحة العالمية أعلى بما يزيد عن ٥٠ بالمائة من الحالات المُسجَّلة فعليًّا.

تشابهت المشكلات الخاصة بتحليلات اتجاهات الأنفلونزا من جوجل والإيبولا في أنَّ خوارزميات التنبؤ المُستخدَمة في كِليهما كانت تعتمد فقط على البيانات الأولية، ولم تأخذ في اعتبارها الظروفَ المتغيرة. افترضَ كلٌّ من هذَين النموذجَين، بصفة أساسية، أن عدد حالات الإصابة سيواصل

الارتفاع بالمعدَّل نفسه في المستقبل مثلما حدثَ قبل بدء التدخل الطبي. ومن الواضح أنه كان يُتوقَع أن تكون للتدابير الطبية وتدابير الصحة العامة تأثيراتٌ إيجابية، ولكنها لم تُضمَّن في النموذج.

سُجِّلَت أولُ إصابة بفيروس زيكا، الذي تنقُله البعوضة الزاعجة، عام ١٩٤٧ في أوغندا، ثم انتشر بعيدًا عن مكان الإصابة الأولى ليصل إلى آسيا والأمريكتَين. أدَّى تفشي فيروس زيكا الحالي، الذي بدأ في البرازيل عام ٢٠١٥، إلى ظهور حالة أخرى من طوارئ الصحة العامة التي تثير قلقًا دوليًّا. كانت ثمَّة دروس مستفادة من العمل الذي قام به مشروع اتجاهات الأنفلونزا من جوجل وخلال تفشي الإيبولا، تتعلَّق بإعداد النماذج الإحصائية باستخدام البيانات الضخمة، وأصبح من المُتفَق عليه عمومًا الآن ضرورة جمع البيانات من مصادر متعدِّدة. ولعلك تتذكَّر أنَّ مشروع اتجاهات الأنفلونزا من جوجل جمعَ البيانات من محرك بحث جوجل فقط.

زلزال نيبال

إذن، ما مستقبل تتبُّع مسار الأوبئة باستخدام البيانات الضخمة؟ استُخدمَت خصائصُ الوقت الحقيقي لسجلات تفاصيل مكالمات الهواتف المحمولة في المساعدة في مراقبة حركة السكان خلال الكوارث، مثلما حدث خلال زلزال نيبال وتفشي أنفلونزا الخنازير في المكسيك. على سبيل المثال، استخدم فريقٌ دوليٌّ تابع لمؤسسة فلومايندر، بالإضافة إلى علماء من جامعتَي ساوثامبتون وأكسفورد، فضلًا عن مؤسساتٍ في الولايات المتحدة والصين، بعد زلزال نيبال الذي وقعَ في الخامس والعشرين من أبريل عام ٢٠١٥، سجلات تفاصيل مكالمات الهواتف المحمولة في تقديم تقديرات لحركة السكان. نسبة كبيرة من سكان نيبال لديهم هواتف محمولة، وباستخدام البيانات المُجَهَّلة لاثنَي عشر مليون مشترك في الخدمة، تمكَّن فريقُ مؤسسة فلومايندر من تتبُّع حركة السكان خلال تسعة أيام من وقوع الزلزال. ترجع هذه الاستجابة السريعة، في جزءٍ منها، إلى وجود اتفاق سارٍ مع مزوِّد الخدمة الرئيسي في دولة نيبال، والذي استُكمِلَت تفاصيله الفنية قبل أسبوع واحد فقط من وقوع الكارثة. ونظرًا لوجود خادم مخصَّص تبلغ السعة التخزينية لقرصه الصلب ٢٠ تيرابايت في مركز بيانات مزوِّدي الخدمة، تمكَّن الفريقُ من بدء العمل على الفور، ما أدَّى إلى إتاحة المعلومات أمام مؤسسات الإغاثة من الكوارث في غضون تسعة أيام فقط من وقوع الزلزال.

البيانات الضخمة والطب الذكي

في كل مرة يزور مريضٌ عيادةَ طبيب أو مستشفًى، تُجمَع بياناتٌ إلكترونية بصفة روتينية. تُشكِّل السجلات الصحية الإلكترونية الوثيقة القانونية لجهات اتصال الرعاية الصحية الخاصة بالمريض؛ ذلك حيث تُسجَّل تفاصيلُ على غرار التاريخ الطبي للمريض، والأدوية الموصوفة، ونتائج الفحوصات. ومن الوارد أيضًا أن تشمل السجلات الصحية الإلكترونية بيانات أجهزة الاستشعار،

مثل فحوصات التصوير بالرنين المغناطيسي. وقد تَجهل البيانات وتَجمَع لأغراضٍ بحثية. كانت هناك تقديرات تشير إلى أنه بحلول عام ٢٠١٥ ستُخزِّن المستشفى العادية في الولايات المتحدة ما يزيد عن ٦٠٠ تيرابايت من البيانات، أغلبها بياناتٌ غير هيكلية. وكان السؤال كيف يمكن التنقيب في هذه البيانات للحصول على معلوماتٍ من شأنها تحسين رعاية المرضى وتقليل التكاليف؟ ما حدث باختصار أننا أخذنا البيانات، سواءٌ الهيكلية أو غير الهيكلية، وحدَّدنا السمات ذات الصلة بمريض أو مجموعة من المرضى، واستخدمنا الأساليبَ الإحصائية على غرار التصنيف والانحدار في إعداد نموذج بالنتائج. تُجمَع الملاحظات الخاصة بحالة المرضى بصفة أساسية بالتنسيق النصي غير الهيكلي، ولتحليل هذه الملاحظات على نحو فعَّال، يتطلَّب الأمر استخدام أساليب معالجة اللغات الطبيعية، كتلك المستخدَمة من قِبل نظام واتسون من شركة آي بي إم، والذي سنتحدَّث عنه في الجزء التالي.

طبقًا لشركة آي بي إم، كان المتوقَّع بحلول عام ٢٠٢٠ أن تتضاعف كمية البيانات الطبية كلَّ ٧٣ يومًا. ومع تزايد استخدامها في مراقبة الأصحاء، أصبحتِ الأجهزة القابلة للارتداء تُستخدَم على نطاق واسع في حساب عدد الخطوات التي نخطوها كلَّ يوم، وقياس احتياجاتنا من السعرات الحرارية وموازنتها، ومتابعة أنماط النوم لدينا، وكذلك تقديم معلومات فورية عن معدل نبضات القلب وضغط الدم. بعد ذلك، تُرفَع المعلومات المُجمَّعة على أجهزة الكمبيوتر وتُحفَظ السجلات على نحو خاص، أو — كما هو الحال أحيانًا — تجري مشاركتها طوعًا مع أصحاب العمل. سيوفِّر هذا التتابع الواقعي للبيانات المتعلِّقة بالأفراد للعاملين في مجال الرعاية الصحية بيانات قيِّمةً عن الصحة العامة، كما سيوفر وسيلةً لملاحظة التغييرات التي تطرأ على الأفراد والتي قد تساعد في تجنُّب الأزمات القلبية، على سبيل المثال. كما أنَّ البيانات المتعلِّقة بفئات السكان ستمكِّن الأطباءَ من تتبُّع الأعراض الجانبية لدواءٍ معين، على سبيل المثال، بناءً على خصائص المرضى.

بعد اكتمال مشروع الجينوم البشري عام ٢٠٠٣، تزايدت أهمية البيانات الوراثية بوصفها جزءًا من السجلات الطبية للأفراد، كما ستُقدِّم ثروةً من البيانات البحثية. كان الهدفُ من مشروع الجينوم البشري وضعَ خريطةٍ بكل الجينات البشرية. يُطلَق على المعلومات الوراثية للكائن الحي مجتمعةً اسم الجينوم. يحتوي الجينوم البشري، إجمالًا، على حوالي ٢٠ ألف جين، ويتطلَّب وضع خريطة لهذا الجينوم نحو ١٠٠جيجابايت من البيانات. ممَّا لا شك فيه أنَّ هذا المجال من أبحاث الوراثة هو مجالٌ شديد التعقيد والتخصُّص والتشعُّب، إلا أن النتائج المترتبة على استخدام أساليب تحليل البيانات الضخمة تسترعي الاهتمام. ومن ثمَّ، حُفِظَت المعلوماتُ التي جُمِعَت عن الجينات في قواعد بياناتٍ ضخمة؛ ولذا، ظهرت مؤخَّرًا مخاوف من احتمالية تعرض هذه المعلومات للقرصنة، ممَّا يؤدِّي إلى تحديد هُويات المرضى الذين ساهموا بحمضهم النووي. وقُدِّم اقتراح بأنه، لأغراضٍ أمنية، يجب إضافة معلوماتٍ زائفة إلى قواعد البيانات، وإنْ كانت ليست بالقدر الذي من شأنه أن يؤثر على الأبحاث الطبية. ازدهرَ مجالُ المعلوماتية الحيوية المتعدِّد التخصُّصات بسبب الحاجة إلى إدارة البيانات الضخمة الناتجة عن علم الجينوم وتحليلها. وتزايدت سرعةُ التسلسل الجيني وقلَّت تكلفتُه كثيرًا خلال السنوات الأخيرة؛ ومن ثمَّ، أصبحَ الآن وضعُ خرائط لجينوم الأفراد أمرًا ممكنًا من الناحية العملية. مع وضع تكاليف ١٥ عامًا من الأبحاث في الاعتبار، بلغت تكلفة تحديد تسلسل الجينوم البشري الأول ما يقارب ٣ ملايين دولار. وبدأت الكثير من الشركات الآن في عرض خدماتها في مجال تحديد تسلسل الجينوم على الأفراد بأسعار معقولة.

تفرَّع من مشروع الجينوم البشري مشروعُ الإنسان الفسيولوجي الافتراضِي الذي يهدف إلى إنشاء عروض تقديمية على أجهزة الكمبيوتر تُتيح للأطباء السريريين محاكاة طُرق العلاج الطبي وتحديد الأنسب منها لكل مريض، وتقوم على البيانات المستقاة من بنك هائل لبيانات مرضى فعليين. وبمقارنة هذه البيانات بأعراض مماثلة أو تفاصيل طبية ذات صلة، يمكن للنموذج المُعَد باستخدام الكمبيوتر أن يتنبَّأ بالنتيجة المرجَّحة التي تتضمَّن طريقة علاج لمريض بعينه. علاوةً على ذلك، تُستخدَم أساليب التنقيب في البيانات التي يمكن دمجها مع عمليات المحاكاة الحاسوبية لإضفاء طابع شخصي على طرق العلاج الطبية حسب كل مريض؛ ومن ثمَّ، يمكن دمج نتائج التصوير بالرنين المغناطيسي في أي محاكاة منها. وهكذا، يُتوقَّع أن يحتوي المريض الرقمي المستقبلي على جميع المعلومات التي تخصُّ مريضًا فعليًّا، والتي تُحدَّث طِبقًا لبيانات الأجهزة الذكية. ولكن، يشكِّل أمن البيانات تحديًا كبيرًا على نحو متزايد أمام المشروع.

استخدام نظام واتسون في الطب

عام ٢٠٠٧، قرَّرت شركة آي بي إم أن تُنشئ جهاز كمبيوتر تتحدَّى به أقوى الشركات المنافسة لها في برنامج المسابقات «جيوباردي»، الذي يُعرَض على شاشة التلفزيون الأمريكي. وُضع واتسون، وهو نظام لتحليل البيانات الضخمة سُمِّي تَيمُّنًا بمؤسِّس شركة آي بي إم، توماس جون واتسون، في مواجهة اثنَين من أبطال برنامج جيوباردي: براد روتر، صاحب سلسلة فوز متتالية بلغت ٧٤ مرة، وكين جينينجز، الذي حصدَ إجمالي مبلغ ٣٫٢٥ ملايين دولار أمريكي. جيوباردي هو برنامج مسابقاتٍ يعطي فيه مضيفُ البرنامج «إجابة»، وعلى المتسابق أن يخمِّن «السؤال». تُجرَى المسابقة بين ثلاثة متسابقين، وتندرج الإجابات أو أدلة الإجابة ضمن عدة فئات على غرار العلوم، والرياضة، وتاريخ العالم إلى جانب فئاتٍ غير مألوفة أو غريبة، مثل «قبل وبعد». على سبيل المثال، إذا كان دليل الإجابة: «يوجد شاهِد قبره في فناء كنيسة هامبشاير ومكتوبٌ عليه: فارس، ووطني، وطبيب، وأديب، ٢٢ مايو ١٨٥٩-٧ يوليو ١٩٣٠»، فستكون الإجابة: «مَن هو السير آرثر كونان دويل؟». وفي الفئة الأقل وضوحًا «اقبضْ على هؤلاء الرجال»، إذا كان دليل الإجابة: «مطلوب القبض عليه في ١٩ جريمة قتل، فَرَّ هذا الرجلُ المولود في بوسطن عام ١٩٩٥، وأُلقِي القبض عليه أخيرًا في سانتا مونيكا عام ٢٠١١»، فستكون الإجابة: «مَن هو وايتي بولجر؟» حُذِفَت أدلة الإجابة، التي قُدِّمَت إلى واتسون في صورة نصوص ورموز صوتية ومرئية، من المسابقة.

تُمثِّل معالجة اللُّغات الطبيعية، كما تُعرَف في مجال الذكاء الاصطناعي، تحديًا كبيرًا لعلوم الكمبيوتر، وكانت ضروريةً لتطوير نظام واتسون. وعطفًا على ما سبق، يجب أن تكون المعلومات قابلةً للوصول إليها واسترجاعها بسهولة، ويُمثِّل هذا الأمر مشكلةً في مجال تعلُّم الآلة. بدأ فريق الأبحاث عمله بتحليل أدلة الإجابة الخاصة بمسابقة جيوباردي طبقًا لنوع الإجابة المعجمي، الذي يُصنِّف نوعَ الإجابة المُحدَّد في الدليل. في المثال الثاني الذي ذكرناه، نوع الإجابة المعجمي هو «المولود في بوسطن». أمَّا المثال الأول، فلا يوجد فيه نوع إجابة معجمي؛ إذ لا تُفيد الضمائرُ هذه العملية كثيرًا. وبتحليل ٢٠ ألف دليل إجابة، عثر فريقُ آي بي إم على ٢٥٠٠ نوع إجابة معجمي

فريد، إلا أن هذا العدد لم يغطِ إلا حوالي نصف أدلة الإجابة فقط. بعد ذلك، يُحلل دليل الإجابة لتحديد الكلمات الرئيسية والعلاقات بينها. وتُسترجَع المستنداتُ ذات الصلة من بيانات الكمبيوتر الهيكلية وغير الهيكلية ويُبحَث فيها. وتُوضَع فرضياتٌ بِناءً على التحليلات المبدئية، وبالبحث في أدلة إجابة أكثر عُمقًا، يُعثَر على الإجابات المُحتمَلة.

للفوز بمسابقة جيوباردي، كان لا بد من استخدام الأساليب السريعة المتطوِّرة فيما يخصُّ معالجة اللغات الطبيعية، وتعلُّم الآلة، والتحليل الإحصائي. وكان من بين العوامل الأخرى الواجب مراعاتها الدقة واختيار الفئة. وأُنشئ معيار للأداء المقبول باستخدام بيانات الفائزين السابقين. وبعد عدة محاولات، جاء الحل في صورة تحليل عميق للأسئلة والأجوبة، أو ما يُسمَّى «ديب كيو إيه»، وهو عبارة عن دمجٍ للكثير من أساليب الذكاء الاصطناعي. يستخدم هذا النظامُ مجموعةً كبيرة من أجهزة الكمبيوتر، التي تعمل بالتوازي ولكنها ليست متصلة بالإنترنت، ويعتمد على الاحتمالية وبراهين الخبراء. بالإضافة إلى التوصل إلى إجابة، يستخدم واتسون خوارزميات حساب حَد الثقة لإتاحة إمكانية العثور على أفضل نتيجة. ولا يُشير واتسون إلى أنه جاهز لإعطاء الإجابة إلا عندما يصل إلى حَد الثقة المُعيَّن، وهو ما يكافئ ضغط المتنافس البشري على زر الجرس. تمكَّن واتسون من هزيمة بَطَلَي جيوباردي. واستُشهِد بمقولة جينينجز، الذي تقبَّل الهزيمة بصدر رحب، حيث قال: «من جانبي، فأنا أرحِّب بسادتنا الجُدد من أجهزة الكمبيوتر».

يسترجع نظام واتسون الطبي، القائم على نظام واتسون الأصلي الخاص بمسابقة جيوباردي، كلًّا من البيانات الهيكلية وغير الهيكلية ويحللها. وبما أنه يبني قاعدة المعارف الخاصة به بنفسه، فإنه بالأساس نظام يُجري نمذجةً لعمليات التفكير البشري في مجال معين. تعتمد التشخيصات الطبية على كل المعلومات الطبية المتوافرة، والتي تكون مُثبَتةً بالأدلة ودقيقةً إلى الحَدِّ الذي تكون معه المُدخلات دقيقةً ومتسقة وتتضمَّن جميع المعلومات ذات الصلة. يتمتَّع الأطباءُ البشريون بالخبرة، ولكنهم غير معصومين من الخطأ، وبعضهم بارعٌ في التشخيص أكثر من غيره. تشبه هذه العملية الآلية المتبَعة في نظام واتسون الخاص بمسابقة جيوباردي، حيثٍ تُؤخَذ في الاعتبار جميع المعلومات ذات الصلة وتُعطَى التشخيصات مع تحديد درجة ثقة لكلٍّ منها. وتسمح تقنيات الذكاء الاصطناعي المُضمَّنة في نظام واتسون بمعالجة البيانات الضخمة، بما في ذلك الكميات الهائلة الناتجة عن التصوير التشخيصي الطبي.

أصبح كمبيوتر واتسون العملاق حاليًّا نظامًا متعدِّد التطبيقات، وحقَّق نجاحًا تجاريًّا هائلًا. علاوةً على ذلك، يشارك واتسون في الجهود الإنسانية، ويحدث هذا — على سبيل المثال — من خلال نظام تحليلاتٍ مفتوح المصدر طُوِّر خصوصًا للمساعدة في تتبُّع انتشار الإيبولا في دولة سيراليون.

خصوصية البيانات الطبية الضخمة

تأكَّدَ بوضوح أن البيانات الضخمة لديها القدرة على التنبؤ بانتشار الأمراض وتخصيص طرق العلاج، ولكن، ماذا عن الوجه الآخر للعُملة: خصوصية البيانات الطبية للأشخاص؟ مع تزايد

استخدام الأجهزة القابلة للارتداء وتطبيقات الهواتف الذكية على وجه الخصوص، طرأت أسئلة على غرار مَن يملك البيانات، وأين تُخزَّن، ومَن يمكنه الوصول إليها واستخدامها، وما مدى تأمينها ضد الهجمات الإلكترونية عبر الإنترنت. ثمَّة الكثير من القضايا الأخلاقية والقانونية التي لن يسعنا تناولها في هذا الكتاب.

قد تصبح البيانات الصادرة من أحد أجهزة متابعة اللياقة البدنية متوافرةً لأحد أصحاب العمل، وتُستخدَم: إمَّا بصورة إيجابية، مثل تقديم علاواتٍ لمَن يستَوفون معايير معينة، وإمَّا بصورة سلبية، مثل تحديد أولئك الذين يُخفقون في تلبية المعايير المطلوبة، الأمر الذي قد يؤدي إلى تسريح العمالة غير المرغوب فيها. في سبتمبر ٢٠١٦، نشر فريقُ أبحاث مشترك، مُكوَّن من علماء من جامعة دارمشتات للتكنولوجيا في ألمانيا وجامعة بادوا في إيطاليا، نتائج دراسة أجرَوها على أمن بيانات أجهزة متابعة اللياقة البدنية. المقلق في الأمر أنه من بين ١٧ جهازًا خضعَ للاختبار، جميعها من مُصنِّعين مختلفين، لم يكن أيٌّ منها مؤمَّنًا بما يكفي لإيقاف التغييرات الجاري إدخالها على البيانات، وأربعة أجهزة فقط هي التي اتخذت إجراءاتٍ للحفاظ على موثوقية البيانات، وتمكَّن أعضاءُ الفريق من تجاوزها جميعًا.

في سبتمبر ٢٠١٦، بعد دورة الألعاب الأولمبية في ريو دي جانيرو، والتي تقرَّر حظر معظم الرياضيين الروس منها بعد تقارير موثَّقة عن برنامج لتعاطي المنشطات تديره الدولة، تعرَّضت السجلات الطبية لرياضيين كبار، من بينهم الشقيقتان ويليامز، وسيمون بايلز، وكريس فروم، للاختراق، وتمَّ الكشف عنها علنًا بواسطة مجموعة من قراصنة الإنترنت الروس على موقع FancyBears.net. لم تكشف هذه السجلات الطبية، التي كانت في حوزة الوكالة العالمية لمكافحة المنشطات (المعروفة ب «وادا») على نظام إدارة البيانات الخاص بها الذي يُدعَى «أدامز» (نظام إدارة وتنظيم مكافحة المنشطات)، سوى استخدامات استثنائية لأغراض علاجية، وعليه فهي لم تُدِن أيًّا من الرياضيين الذين تعرَّضوا للتنمُّر الإلكتروني. ومن المرجَّح أن الاختراق الأولي لنظام إدارة وتنظيم مكافحة المنشطات تمَّ بواسطة حسابات رسائل البريد الإلكتروني للتصيُّد المُوجَّه. يُستخدَم هذا الأسلوب، الذي يبدو فيه أن رسالةً إلكترونية مُرسلة من مصدر كبير موثوق داخل المؤسسة، مثل مقدِّم خدمة الرعاية الصحية، إلى عضو أحدث من المؤسسة ذاتها، للحصول على نحو غير قانوني على معلوماتٍ حسَّاسة على غرار كلمات المرور وأرقام الحسابات عن طريق برنامج ضار يتم تنزيله.

أصبحَ تحصين قواعد البيانات الطبية الضخمة ضد الهجمات الإلكترونية، وما يترتَّب عليه من ضمان خصوصية المرضى، هاجسًا متناميًا. يجوز قانونًا بيعُ البيانات الطبية الشخصية المُجهَّلة، ولكن من الممكن في بعض الأحيان تحديد هُويَّات المرضى. في ممارسة قيمة تهدف إلى الكشف عن الثغرات الأمنية في البيانات التي من المفترض أن تكون آمنة، تمكَّنت عالمتان من مختبر هارفارد لخصوصية البيانات، هما لاتانيا سويني وجي سو يو، باستخدام بياناتٍ طبية «مُشفَّرة» (أي إنها مختلطة ومشوَّشة حتى لا يمكن قراءتها بسهولة، انظر الفصل السابع)، متاحة بصفة قانونية، ومنشأها كوريا الجنوبية، من فك تشفير معرِّفاتٍ فريدة في السجلات، وتحديد هُويات المرضى من خلال مقارنتها بالسجلات العامة.

تَعد السجلات الطبية بالغة القيمة لدى المجرمين الإلكترونيين. عام ٢٠١٥، أعلنت شركة أنثيم للتأمين الصحي أن قواعد بياناتها قد تعرَّضت للاختراق، ما أثَّر على بيانات أكثر من ٧٠ مليون شخص. تعرَّضت بياناتٌ مهمة لتحديد هُويات الأشخاص، مثل الاسم، والعنوان، ورقم التأمين الاجتماعي، للاختراق على يد ديب باندا، وهو فريق صيني من المخترقين الإلكترونيين، باستخدام كلمات مرور مسروقة للوصول إلى النظام وتحميل برنامج ضار من نوع حصان طروادة. الخطير في الأمر أن أرقام التأمين الاجتماعي، أحد المعرِّفات الفريدة من نوعها في الولايات المتحدة الأمريكية، لم تكن مُشفَّرة، الأمر الذي ترك مجالًا واسعًا لاحتمالية سرقة الهُويات. تبدأ الكثير من الاختراقات الأمنية بأخطاء بشرية: مثل الانشغال وعدم ملاحظة التغيُّرات الطفيفة في محدِّدات مواقع الويب «يو آر إل»، وفقدان أجهزة على غرار محركات الأقراص المحمولة أو سرقتها، أو حتى في بعض الأحيان إحلالها بأخرى تحتوي على برامج ضارة تُحمَّل على الفور بمجرد أن يضع موظفٌ غير مرتاب الجهاز في منفذ «يو إس بي». ويكون كذلك الموظَّفون المستاءون، وأخطاء الموظَّفين غير المقصودة، هي المتهم الرئيسي فيما يقع من تسريباتٍ للبيانات لا حصرَ لها.

بدأت التحفيزاتُ الجديدة لاستخدام البيانات الضخمة في مجال إدارة الرعاية الصحية تُطلَق بمعدَّل متزايد من قِبَل مؤسساتٍ ذات شهرة عالمية على غرار مجموعة مايو كلينيك، ومجموعة جونز هوبكنز الطبية في الولايات المتحدة الأمريكية، وهيئة الخدمات الصحية الوطنية في المملكة المتحدة، ومستشفى جامعة كليرمون فيران في فرنسا. منحت الأنظمة المستندة إلى الحوسبة السحابية المستخدمين المُصرَّح لهم بإمكانية الوصول إلى البيانات من أي مكان في العالم. وإذا ذكرنا مثالًا واحدًا على ذلك، فسيكون خطط هيئة الخدمات الصحية الوطنية لإتاحة سجلات المرضى عبر الهواتف المحمولة بحلول عام ٢٠١٨. وكان من شأن هذه التطورات أن تتسبَّب، لا محالة، في المزيد من الهجمات على البيانات التي تستخدمها، مع إدراك ضرورة بذل جهود كبيرة لتطوير أساليب أمان فعَّالة لضمان سلامة هذه البيانات.

الفصل السادس

البيانات الضخمة والشركات الكبرى

في عشرينيات القرن العشرين، وظَّفت شركة جيه ليونز وشركائه، وهي شركة بريطانية تمتلك سلسلة مطاعم وشركات أغذية وفنادق، تشتهر بسلسلة مقاهي «كورنر هاوس»، عالِمَ رياضياتٍ شاب من جامعة كامبريدج يُدعى جون سيمونز، لتولي أعمال الإحصاء. عام ١٩٤٧، أرسِل كلٌّ من رايموند تومسون وأوليفر ستاندينجفورد، اللذَين عيَّنهما سيمونز، في زيارة إلى الولايات المتحدة الأمريكية لتقصي الحقائق. وخلال هذه الزيارة، تعرَّفا للمرة الأولى على أجهزة الكمبيوتر الإلكترونية وقدرتها على إجراء العمليات الحسابية الروتينية. وسعى سيمونز، منبهرًا بنتائج الزيارة، لإقناع ليونز بشراء جهاز كمبيوتر.

أثمر التعاونُ مع موريس ويلكس، الذي كان منخرطًا في ذلك الوقت في تصميم الكمبيوتر الآلي لتخزين التأخير الإلكتروني في جامعة كامبريدج، عن كمبيوتر ليونز المكتبي الإلكتروني. كان هذا الكمبيوتر يعمل باستخدام البطاقات المُثقَّبة، واستخدم لأول مرة عام ١٩٥١ من قِبَل شركة ليونز في إجراء العمليات الحسابية الأساسية، مثل جمع أعمدة تحتوي على أرقام. وبحلول عام ١٩٥٤، أسَّست شركة ليونز شركتها الخاصة لأجهزة الكمبيوتر، وبدأت في تصميم كمبيوتر ليونز المكتبي الإلكتروني الثاني، وتبعَه كمبيوتر ليونز المكتبي الإلكتروني الثالث. وعلى الرغم من أن بداية العمل بأجهزة الكمبيوتر المكتبية الأولى جاءت في وقت مبكر، في خمسينيات القرن العشرين، فإن هذه الآلات الأولى لم تكن موثوقة، وكانت تطبيقاتها محدودة؛ بسبب استخدامها للصمامات (٦ آلاف صمام في حالة كمبيوتر ليونز المكتبي الإلكتروني الأول) والشرائط الممغنطة، وسعة التخزين المحدودة للغاية لذاكرة الوصول العشوائي. اشتهر كمبيوتر ليونز المكتبي الإلكتروني الأول على نطاق واسع بأنه كمبيوتر إدارة الأعمال الأول، الأمر الذي مهَّد الطريقَ أمام التجارة الإلكترونية الحديثة، وبعد عدة عمليات دمج مع شركاتٍ أخرى، أصبح في نهاية المطاف جزءًا من شركة إنترناشونال كمبيوترز ليمتد حديثة التكوين عام ١٩٦٨.

التجارة الإلكترونية

لم تكن أجهزة كمبيوتر ليونز المكتبية الإلكترونية، وأجهزة الكمبيوتر المركزية الضخمة التي تلتها، تصلح إلا لمهام معالجة الأرقام التي تُعد جزءًا من مهام المحاسبة والمراجعة. وأصبح الموظَّفون، الذين كانوا يقضون أوقاتهم سابقًا في حساب أعمدة من الأرقام، يقضون أوقاتهم في إعداد البطاقات المُثقَّبة، وهي مهمة لا تقل عن سابقتها مَلَلًا، بل وتحتاج إلى تحري المستوى نفسه من الدقة الفائقة.

منذ أن أصبح استخدام أجهزة الكمبيوتر مجديًا بالنسبة إلى المؤسسات التجارية، ظهر الاهتمام بكيفية استخدامها في رفع الكفاءة، وتقليل التكاليف، وتحقيق الأرباح. وأدَّى تصميم الترانزستور واستخدامه

في أجهزة الكمبيوتر المتاحة تجاريًّا إلى صُنع أجهزة أصغر حجمًا من ذي قبل، وفي أوائل السبعينيات من القرن العشرين، ظهرت أولى أجهزة الكمبيوتر الشخصية. ولكن، لم تُطرَح هذه الفكرة تجاريًّا حتى عام ١٩٨١ عندما طرحت شركة إنترناشونال بيزنس ماشينز (آي بي إم) كمبيوتر آي بي إم الشخصي في الأسواق، مع استخدام الأقراص المرنة في تخزين البيانات. وكانت إمكانات معالجة النصوص وجداول البيانات التي امتلكتها الأجيال اللاحقة من أجهزة الكمبيوتر الشخصية مسئولةً إلى حَدٍّ كبير عن تخفيف الكثير من أعباء الأعمال المكتبية الروتينية.

وعلى ضوء التقنية التي أتاحت الإمكانية لتخزين البيانات إلكترونيًّا على أقراص مرنة، سرعان ما ظهرت فكرة أنَّ المؤسسات قد تُدار بفاعلية في المستقبل دون استخدام الورق. في ١٩٧٥، تنبأ مقالٌ نُشِر في مجلة بيزنس ويك الأمريكية بأن أماكن العمل الخالية تقريبًا من الورق يمكن أن تُصبح واقعًا بحلول عام ١٩٩٠. واقترحَ المقال أنه من خلال الاستغناء عن استخدام الورق أو تقليله إلى حَدٍّ كبير، قد يُصبح مكان العمل أكثر فاعليةً وقد تقل التكاليف. تراجَع مُعدَّل استخدام الورق في أماكن العمل لفترة من الوقت خلال ثمانينيات القرن العشرين عندما نُقِلت كثيرٌ من الأعمال الورقية التي كان من المعتاد رؤيتها في خزائن الملفات إلى أجهزة الكمبيوتر، ثم سجَّل هذا الاستخدام أعلى معدلاته على الإطلاق عام ٢٠٠٧، وكانت النسخ المصوَّرة هي المسئولة عن السواد الأعظم من هذه الزيادة. منذ عام ٢٠٠٧، ظلَّ استخدام الورق يتراجع تدريجيًّا، ويرجع الفضل الأكبر في ذلك إلى زيادة استخدام الهواتف الذكية وتسهيلاتٍ على غرار التوقيع الإلكتروني.

على الرغم من أن التطلُّعات المتفائلة التي ظهرت منذ بداية العصر الرقمي إلى جعل أماكن العمل خاليةً من الورق لم تتحقَّق وقتها، حدثت ثورةٌ في بيئة العمل بفِعل البريد الإلكتروني، وبرامج معالجة النصوص، وجداول البيانات الإلكترونية. إلا أن استخدام الإنترنت على نطاق واسع هو ما جعلَ التجارة الإلكترونية مُقترَحًا عمليًّا.

لعلَّ التسوُّق عبر الإنترنت هو المثالُ الأشهر. فنحن، باعتبارنا عملاء، نستمتع برفاهية التسوق من المنزل وتجنُّب الطوابير التي تستغرق وقتًا طويلًا. السلبيات التي يتعرَّض لها العملاء قليلة، ولكن، بِناءً على نوع المعاملة، قد تؤدِّي عدم القدرة على التواصل وجهًا لوجه مع موظَّفي المتاجر إلى تجنُّب استخدام الشراء عبر الإنترنت. وعلى نحو متزايد، أصبح من الممكن التغلُّب على هذه المشكلات من خلال تسهيلات تقديم المشورة للعملاء عبر الإنترنت مثل «الدردشة الفورية»، والتقييمات عبر الإنترنت، والتصنيف بالنجوم، بالإضافة إلى مجموعة اختياراتٍ ضخمة من السلع والخدمات فضلًا عن سياسات الإرجاع السخية. بالإضافة إلى شراء السلع ودفع مقابلها، أصبح بالإمكان حاليًّا دفع الفواتير، وإجراء المعاملات المصرفية، وشراء تذاكر الطيران، والوصول إلى مجموعة من الخدمات الأخرى جميعها عبر الإنترنت.

يعمل موقع إيباي بأسلوب مختلف نوعًا ما، ويستحق أن يُذكَر نظرًا لكمية البيانات الهائلة التي يُنتجها. بالنظر إلى المعاملات التي تُجرَى عبر عمليات البيع وعطاءات المزادات، ينتج إيباي حوالي ٥٠تيرابايت من البيانات يوميًّا، وتُجمَع هذه البيانات من كل عملية بحث، وبيع، ومزاد تُجرَى على الموقع بواسطة مُستخدِميه النُّشطاء الذين يُزعَم أن عددَهم ١٦٠ مليون مُستخدِم من ١٩٠ دولة.

باستخدام هذه البيانات وأساليب التحليل المناسبة، تمكن الموقع حاليًّا من تنفيذ أنظمة توصية مثيلة لأنظمة نتفليكس، والتي سنتحدَّث عنها لاحقًا في هذا الفصل.

تُوفِّر مواقع شبكات التواصل الاجتماعي للشركات ملاحظاتٍ فوريةً عن كل شيءٍ من الفنادق والعطلات إلى الملابس، وأجهزة الكمبيوتر، والزبادي. باستخدام هذه المعلومات، يمكن للشركات معرفة العناصر التي تحقِّق نجاحًا، وحجم هذا النجاح، والجوانب المثيرة للشكاوى، مع حَلِّ المشكلات قبل أن تخرج عن نطاق السيطرة. بل إن القيمة الأكبر لهذه المعلومات هي منح القدرة على التنبؤ بما يرغب العملاء في شرائه بناءً على عمليات الشراء السابقة أو نشاط العملاء على الموقع الإلكتروني. تجمع مواقع شبكات التواصل الاجتماعي، مثل فيسبوك وتويتر، كمياتٍ هائلةً من البيانات غير الهيكلية التي يمكن أن تستفيد الشركات بها تجاريًّا في حال استخدام أساليب التحليل المناسبة. كما تشارك مواقع السياحة والسفر، مثل تريب أدفايزر، المعلومات مع جهاتٍ أخرى.

إعلانات الدفع مقابل النقر

أصبح الخبراء يُقرُّون الآن، على نحوٍ متزايد، بأنَّ الاستخدام الصحيح للبيانات الضخمة من شأنه أن يوفِّر بياناتٍ مفيدةً ويجتذب عملاء جُددًا عبر الترويج المُحسَّن للسلع واستخدام دعاية موجَّهة على نحو أفضل. في كل مرة نستخدم الويب، نشاهِد إعلانات عبر الإنترنت لا محالة، بل وقد ننشر بأنفسنا إعلاناتٍ مجانيةً على العديد من مواقع المزادات على غرار إيباي.

إن أحد أشهر أنواع الإعلان هو ذلك الذي يتبع نموذج الدفع مقابل النقر، وهو نظام تظهر خلاله إعلاناتٌ ذات صلة عند إجراء عملية بحثٍ عبر الإنترنت. إذا أرادت شركة أن تُعرَض إعلاناتها عند الاستعلام عن مصطلح بحث معين، فإنها تضع عطاءً مع مزوِّد الخدمة على كلمة رئيسية تتعلَّق بمصطلح البحث هذا. كما أنها تُعلن ميزانيةً يومية قصوى. وتُعرَض الإعلانات بالترتيب وفقًا لنظام يستند جزئيًّا إلى أيِّ المُعلِنين قدَّم العطاءَ الأعلى على هذا المصطلح.

إذا نقرتَ فوق إعلانٍ لأحد المُعلِنين، فسيكون عليه أن يدفع إلى مزوِّد الخدمة قيمة العطاء الذي حدَّده. ولا تدفع الشركات المال إلا إذا نقرَ طرفٌ مهتمٌّ فوق إعلاناتها؛ ومن ثَمَّ، يجب أن تكون هذه الإعلانات ملائمةً تمامًا لمصطلح البحث حتى تزداد أرجحية أن ينقر متصفِّحو الويب فوقها. وتضمن خوارزمياتٌ دقيقة أن يُحقِّق مزوِّد الخدمة، مثل جوجل أو ياهو، أقصى عائدٍ ممكن. ويُعَد جوجل آدووردز (المعروف الآن بإعلانات جوجل أو جوجل آدز) أفضل تطبيق معروف لإعلانات الدفع مقابل النقر. عندما نُجري بحثًا على محرك بحث جوجل، ينشئ آدووردز الإعلانات التي تظهر تلقائيًّا على جانب الشاشة بواسطة آدووردز. الجانب السلبي في هذا النموذج هو أن النقرات قد تكون باهظة، كما أن هناك حدًّا لعدد الأحرف المسموح باستخدامها حتى لا يشغل الإعلان حيزًا أكثر من اللازم.

يمثِّل النقر الاحتيالي مشكلةً أيضًا. على سبيل المثال، قد تنقر شركة منافسة فوق إعلانك بصورة مُتكرِّرة حتى تستنفد ميزانيتك اليومية. أو يمكن استخدام برنامج ضار، يُسمى كليكبوت، لإنتاج نقراتٍ زائفة. والمُعلِن وحده هو مَن يقع ضحيةً لهذا النوع من الاحتيال؛ لأن مزوِّد الخدمة يحصل على أمواله دون مشاركة أي عميل. ولكن، بما أن ضمان الأمن؛ ومن ثمَّ حماية المشروع التجاري المربح، يصبُّ في مصلحة مزوِّدي الخدمة، تُبذَل جهودٌ بحثية كبيرة من أجل مكافحة الاحتيال. ربما كانت أبسط الطرق هي متابعة عدد النقرات المطلوبة في المتوسط لإتمام عمليات الشراء. وإذا حدثت زيادة مفاجئة في عدد النقرات أو تنفيذ عددٍ كبير من النقرات دون إجراء عمليات شراءٍ فعلية، فمن المرجَّح أن يكون هذا نقرًا احتياليًّا.

على النقيض من ترتيبات الدفع مقابل النقر، من الجليِّ أن الإعلانات المُوجَّهة تعتمد على سجل نشاط كلِّ شخص على الإنترنت. ولكي نعرف كيفية عمل هذا النوع من الإعلانات، سنبدأ بتناول ملفات تعريف الارتباط، التي لم أسهب في الحديث عنها في الفصل الأول، بمزيدٍ من التفصيل.

ملفات تعريف الارتباط

ظهر هذا المصطلح للمرة الأولى عام ١٩٧٩ عندما تضمَّن نظامُ التشغيل يونيكس برنامجًا يُسمَّى «فورتشن كوكي»، والذي كان يُرسِل عروضَ أسعار عشوائيةً إلى المستخدمين مستخرجةً من قاعدة بياناتٍ ضخمة. لملفات تعريف الارتباط العديدُ من الأشكال، وتنشأ جميعها خارجيًّا وتُستخدَم في متابعة نشاطٍ ما على أحد المواقع الإلكترونية أو أجهزة الكمبيوتر. عندما تزور موقعًا إلكترونيًّا، يرسل خادم ويب رسالةً إلى متصفحك، وهذه الرسالة عبارة عن ملف صغير يُخزَّن على جهاز الكمبيوتر لديك. تُعد هذه الرسالة أحدَ الأمثلة على ملفات تعريف الارتباط، إلا أن ثمَّة الكثير من الأنواع الأخرى، مثل تلك التي تُستخدَم لأغراض مصادقة المستخدم، وتلك المستخدمة في تعقُّب الجهات الخارجية.

الإعلانات المُوجَّهة

تُجمَع بيانات كلِّ نقرة تنقرها على الإنترنت وتُستخدَم في الإعلانات الموجَّهة.

تُرسَل هذه البيانات إلى شبكات إعلانية لجهاتٍ أخرى، وتُخزَّن على جهاز الكمبيوتر لديك في صورة ملف تعريف ارتباط. وعندما تنقر على مواقع أخرى تدعمها هذه الشبكات، ستُعرَض إعلاناتٌ عن منتجاتٍ عاينتَها سابقًا على شاشتك. باستخدام لايتبيم، أحد البرامج الإضافية المجانية لمتصفح موزيلا فايرفوكس، يمكنك أن تتبع مسار الشركات التي تجمع بيانات نشاطك على الإنترنت.

أنظمة التوصية

توفِّر أنظمة التوصية أو الاقتراح آلية تصفية تُزوِّد المستخدِمين بمعلوماتٍ بِناءً على اهتماماتهم. تعرض أنواعٌ أخرى من أنظمة التوصية، لا تعتمد على اهتمامات المستخدمين، ما يتصفَّحه العملاءُ الآخرون في الوقت الحقيقي، وعادةً ما تظهر هذه التوصيات على أنها «الأكثر تداولًا». ومن أمثلة الشركات التي تستخدم هذه الأنظمة نتفليكس، وأمازون، وفيسبوك.

ثمَّة طريقة لتحديد المنتجات التي يُوصَّى بها للعملاء وهي «التصفية التعاونية». بوجه عام، تستخدم الخوارزميةُ البياناتَ التي تُجمَع عن كل عميل على حدة من عمليات الشراء والبحث السابقة التي أجراها، وتُقارِن هذه البيانات بقاعدة بياناتٍ ضخمة تتضمَّن العناصر التي نالت استحسان العملاء الآخرين وتلكَ التي لم تَنَل استحسانهم؛ وذلك من أجل تقديم توصياتٍ مناسبة بشأن عمليات الشراء الجديدة. ولكن، لا تؤدي المقارنة البسيطة بوجهٍ عام إلى نتائج جيدة. دعونا نتناول المثال الآتي.

لنفترض أن مكتبةً عبر الإنترنت تبيع كتابَ طبخٍ إلى أحد العملاء. قد يكون من السهل بالتالي أن تُوصي العميلَ بجميع كتب الطبخ، ولكن من غيرِ المُرجَّح أن ينجح هذا في ضمان عمليات شراءٍ جديدة. فثمَّة الكثير جدًّا من كتب الطبخ، والعميل على دراية بالفعل أنه يهوي كتبَ الطبخ. ما نحتاج إليه في هذه الحالة هو طريقة لتقليل عدد الكتب المُوصَى بها ليكون مقصورًا فقط على الكتب التي من المُحتمَل أن يشتريها العميل. دعونا نُلقي نظرةً على ثلاثة عملاء هم سميث، وجونز، وبراون، إلى جانب مشترياتهم من الكتب (جدول ٦-١).

جدول ٦-١: الكتب التي اشتراها كل من سميث، وجونز، وبراون

	فن إعداد السلطة	الباستا اليوم	مستقبل الحلويات	عصائر ومشروبات
سميث	تمَّ الشراء		تمَّ الشراء	
جونز	تمَّ الشراء			تمَّ الشراء
براون		تمَّ الشراء	تمَّ الشراء	تمَّ الشراء

السؤال الذي يحاول نظام التوصية الإجابة عنه هو: ما الكتب التي يجدر التوصية بها إلى سميث وأيها إلى جونز؟ نريد أن نعرف ما إذا كان سميث من المُرجَّح أن يشتري كتابَ «الباستا اليوم» أم كتاب «عصائر ومشروبات».

ولكي نفعل هذا، علينا أن نستخدم طريقةً إحصائية لطالما استخدمناها في مقارنة المجموعات وتُسمَّى «معامل تشابه جاكار». ويُعرَّف بأنه عدد العناصر المشتركة بين مجموعتَين مقسومًا على إجمالي عدد العناصر المختلفة في المجموعتَين. ويقيس معامل التشابه التماثُل بين المجموعتَين على أنه نسبة العناصر المشتركة بينهما. وتُعرَّف مسافة جاكار بأنها واحد ناقص معامل تشابه جاكار، وتقيس عدم التماثُل بين المجموعتَين.

بالنظر مرةً أخرى إلى جدول ٦-١، نرى أن سميث وجونز اشتريا الكتاب نفسه، «فن إعداد السلطة». وبالمقارنة بينهما نرى أنهما اشتريا ثلاثة كتب مختلفة؛ «فن إعداد السلطة»، و«مستقبل الحلويات»، و«عصائر ومشروبات». وهذا يعطيهما معامل تشابه جاكار يساوي ٣/١، ومسافة جاكار تساوي ٣/٢. يوضِّح جدول ٦-٢ العمليات الحسابية الخاصة جميع الأزواج المحتملة من العملاء.

جدول ٦-٢: معامل تشابه جاكار ومسافة جاكار

	عدد الكتب المشتركة	إجمالي عدد الكتب المختلفة المُشتراة	معامل تشابه جاكار	مسافة جاكار
سميث وجونز	١	٣	٣/١	٣/٢
سميث وبراون	١	٤	٤/١	٤/٣
جونز وبراون	١	٤	٤/١	٤/٣

يسجِّل سميث وجونز معاملَ تشابه جاكار أعلى، أو درجة تماثُل أعلى، من سميث وبراون. وهذا يعني أن العادات الشرائية لدى سميث وجونز متقاربة، ومن ثَمَّ، نوصي بكتاب «عصائرِ ومشروبات» إلى سميث. ما الذي يجدر بنا التوصية به إلى جونز؟ يسجِّل سميث وجونز معاملَ تشابه جاكار أعلى من جونز وبراون، وعليه، فإننا نوصي بكتاب «مستقبل الحلويات» إلى جونز.

والآن، لنفترض أن العملاء يُقيِّمون عمليات الشراء باستخدامِ نظام تقييم بخمس نجوم. للاستفادة بهذه المعلومات وتوظيفها، علينا أن نعثر على عملاء آخرين أعطَوا التقييمَ نفسه لكتبٍ بعينها، والاطلاع على مشترياتهم الأخرى مع أخذ تاريخهم الشرائي في الاعتبار. التقييم بالنجوم لكل عملية شراءٍ مُوضَّحٌ في جدول ٦-٣.

جدول ٦-٣: التقييم بالنجوم لكل المشتريات

	فن إعداد السلطة	الباستا اليوم	مستقبل الحلويات	عصائر ومشروبات
سميث	٥		٣	
جونز	٢			٥
براون		١	٤	٣

في هذا المثال، تَوصَف طريقة حساب مختلفة، تَسمَّى «مقياس تشابه جَيب التمام»، ويُراعى فيها نظام التقييم باستخدام النجوم. في طريقة الحساب هذه، تُمثِّل المعلومات المُعطاة في جدول التقييم بالنجوم في صورة متجهات. ويُحدَّد عادةً طول المتجه أو مقداره بالواحد الصحيح، ولا يؤدِّي طول المتجه أو مقداره دورًا آخر في الحسابات. يُستخدَم اتجاه المتجه باعتباره وسيلةً لتحديد مدى تماثُل متجهَين؛ ومن ثَمَّ، الجانب صاحب أفضل تقييم بالنجوم. بناءً على نظرية فضاء المتجهات، يتم إيجاد قيمة لتشابه جَيب التمام بين المتجهَين. وتختلف طريقة الحساب هذه نوعًا ما عن طريقة حساب المثلثات المألوفة، إلا أن الخصائص الأساسية تظل قائمةً بأن يأخذ جيب التمام قيمًا تتراوح ما بين صفر وواحد. على سبيل المثال، إذا وجدنا أن تشابه جَيب التمام بين متجهَين، يُمثِّل كلٌّ منهما تقييمَ نجوم أحدِ الأشخاص، يساوي واحدًا، فسيكون قياس الزاوية بينهما صفرًا؛ وذلك لأن جيب التمام لصفر يساوي واحدًا؛ ومن ثَمَّ، لا بد أنهما منطبقان ويمكننا أن نستنتج أن الشخصَين متماثلان في الذوق. وكلما زادت قيمة تشابه جَيب التمام، زادَ حجم هذا التماثُل.

إذا أردت الاستزادة بمعرفة التفاصيل الرياضية، فيمكنك مطالعة المراجع الموجودة في جزء «قراءات إضافية» في نهاية الكتاب. المثيرُ للاهتمام من وجهة نظرنا أن تشابه جَيب التمام بين سميث وجونز يبلغ ٠٫٣٥٠، وبين سميث وبراون يبلغ ٠٫٤٠٤. وهذا الناتج على النقيض من الناتج السابق؛ ما يعني أن سميث وبراون متقاربان في ذوقيهما أكثر من سميث وجونز. بعبارة أخرى، يمكن تفسير ذلك بأن رأي كلٍّ من سميث وبراون في كتاب «مستقبل الحلويات» متقاربان أكثر من رأي سميث وجونز في كتاب «فن إعداد السلطة».

تستخدم خوارزميات التصفية التعاونية من قِبَل كلٍّ من شركتَي نتفليكس وأمازون، اللتَين سنتحدَّث عنهما في الجزء التالي مباشرة.

أمازون

في ١٩٩٤، أسَّسَ جيف بيزوس شركة كادابرا، وسرعان ما غيَّر اسمها إلى أمازون، وفي ١٩٩٥ أُطلق موقع Amazon.com. كانت الشركة في الأساس عبارةً عن مكتبة عبر الإنترنت، وأصبحت حاليًّا شركةً تجارة إلكترونية دولية يبلغ عدد عملائها ٣٠٤ ملايين عميل من جميع أنحاء العالَم. تعمل الشركة في مجال إنتاج وبيع مجموعة متنوِّعة من السلع، بدايةً بالأجهزة الإلكترونية وانتهاءً بالكتب، وحتى الأغذية الطازجة مثل الزبادي، والحليب، والبيض عبر متجر «أمازون فريش». كما أنها شركة رائدة في مجال البيانات الضخمة؛ حيث تقدِّم خدمات أمازون ويب للشركات حلولَ بياناتٍ ضخمة تستند إلى الحوسبة السحابية، باستخدام أدوات متطوِّرة تعتمد على نظام هادوب.

جمعت أمازون بياناتٍ عن الكتب المُشتراة، والكتب التي عاينها العملاء ولكنهم لم يشتروها، والفترة التي قضَوها في البحث عن كتابٍ معيَّن، وما إذا اشتروا الكتب التي حفظوها في قائمة التفضيلات أم لا. وباستخدام هذه البيانات، تمكَّنت أمازون من تحديد المبالغ التي أنفقها العملاء على الكتب شهريًّا

أو سنويًّا، وتحديد ما إذا كانوا عملاء معتادين أم لا. في بداية نشأة الشركة، كانت البيانات التي جمعتها أمازون تُحلَّل باستخدام الأساليب الإحصائية التقليدية. فكانت تُؤخَذ عينات عن الأشخاص، وبِناءً على أوجه التماثل التي يتم إيجادها، قد تعرض أمازون المزيدَ من العناصر المشابهة على العملاء. ولتحسين هذا الأسلوب على نحو أفضل، تقدَّم باحثون من شركة أمازون، عام ٢٠٠١، بطلبٍ للحصول على براءة اختراع لأسلوب يُسمَّى «التصفية التعاونية القائمة على العنصر»، ونالَ الطلبُ الموافقة. يبحث هذا الأسلوب عن العناصر المتشابهة، وليس العملاء المتشابهين.

تجمع أمازون كمياتٍ هائلةً من البيانات، بما في ذلك العناوين، ومعلومات الدفع، وتفاصيل كلِّ ما تصفَّحه الشخصُ أو اشتراه منهم. وتستخدم أمازون بياناتها في تشجيع العملاء على إنفاق المزيد من المال على سلعها، وذلك من خلال محاولة إجراء أكبر قدر ممكن من أبحاث السوق عن العملاء. في حالة الكتب، على سبيل المثال، لا تحتاج أمازون إلى توفير مجموعة ضخمة من الكتب فحسب، بل تحتاج أيضًا إلى تركيز توصياتها على كل عميل على حدة. فإذا اشتركتَ في خدمة أمازون برايم، فستتبع الشركة أيضًا الأفلام التي تشاهدها وعاداتك في القراءة. يستخدم الكثير من العملاء الهواتف الذكية التي تحتوي على خاصية نظام تحديد المواقع العالمي (جي بي إس)، الأمر الذي يُمَكِّن أمازون من جمع البيانات التي توضِّح الوقت والموقع. ويُستخدَم هذا الكم الهائل من البيانات في إنشاء ملفات تعريفٍ للعملاء تتيح مطابقة الأفراد المتشابهين بتوصياتهم.

منذ ٢٠١٣، بدأت أمازون في بيع بيانات تعريف العملاءِ إلى المُعلِنين من أجل الترويج لخدمات ويب الخاصة بها، ما نتجَ عنه نمو كبير للشركة. وفيما يتعلَّق بخدمات أمازون ويب، منصة الشركة للحوسبة السحابية، يُعَد الأمان أمرًا شديدَ الأهمية ومتعدِّد الأوجه. وما كلمات المرور، وأزواج المفاتيح، والتوقيعات الرقمية إلا مجرد أمثلة قليلة على أساليب الأمان المعمول بها لضمان أن تكون حسابات العملاء متاحةً فقط لأولئك الذين يملكون بيانات المصادقة الصحيحة.

تحظى بيانات أمازون بالمستوى نفسه من الحماية المتعدِّدة والتشفير باستخدام خوارزمية «إيه إي إس» (معيار التشفير المتقدِّم) من أجل تخزينها في مراكز البيانات المُخصَّصة لها في جميع أنحاء العالم، و«إس إس إل» (بروتوكول طبقة المنافذ الآمنة)، المعيار الصناعي، في إنشاء وصلة آمنة بين جهازَين، مثل إنشاء رابط بين الكمبيوتر المنزلي وموقع Amazon.com.

أمازون هي الشركة الرائدة في مجال «الشحن الاستباقي» بِناءً على أساليب تحليل البيانات الضخمة. تدور الفكرة حول استخدام البيانات الضخمة في توقُّع السلع التي قد يطلبها العملاء. وتدور الفكرة الأصلية حول شحن المنتجات إلى مركز التوزيع قبل إجراء الطلب فعليًّا. وكإضافة بسيطة، يمكن شحن المنتج إلى العميل مع تقديم مفاجئة مجانية له في حال استحسانه للمنتج. وعلى ضوء سياسة أمازون للاسترجاع، لا تُعد هذه فكرةً سيئة. كان من المتوقَّع أن أغلب العملاء سيحتفظون بالمنتج الذي طلبوه بما أنه كان يعتمد على تفضيلاتهم الشخصية، التي توصَّلت إليها الشركة باستخدام أساليب تحليل البيانات الضخمة. توضِّح أيضًا براءة اختراع الشحن الاستباقي، التي حصلت عليها أمازون عام ٢٠١٤، أيضًا أن رضا العملاء يمكن شراؤه بإرسال هدية ترويجية. إنَّ رضا العملاء، وزيادة المبيعات عبر التسويق المُوجَّه، وتقليل زمن التوصيل، جميعها أمورٌ تجعل أمازون تؤمن بأنها شركة جديرة بالاهتمام. تقدَّمت أمازون أيضًا بطلبٍ للحصول على براءة اختراع

التوصيل بالطائرات بدون طيَّار، وأسمته برايم إير. في سبتمبر ٢٠١٦، خففَت إدارة الطيران الفيدرالية الأمريكية من قوانين تشغيل الطائرات بدون طيَّار من قِبَل المؤسسات التجارية، ما سمحَ لها، في ظروفٍ خاصة تخضع للرقابة الشديدة، بأن تطير خارج مجال رؤية مَن يتحكَّم بها. ربما كانت هذه الخطوة الأولى في مسعى أمازون إلى توصيل الشحنات خلال ٣٠ دقيقة من طلبها، وربما أدَّى هذا إلى توصيل الحليب بالطائرات بدون طيَّار بعدما تكشف أجهزة الاستشعار في ثلاجتك الذكية أنَّ ما بها من حليب أوشكَ أن ينتهي.

أمازون جو، متجر مواد غذائية موجود في سِيَاتل، وهو الأول من نوعه الذي لا يتطلَّب منك الدفع لدى أمين خزينة قبل الخروج من المتجر. حتى ديسمبر ٢٠١٦، كان المتجر متاحًا فقط لموظفي أمازون، وتأجَّلت الخطط لأن يصبح متاحًا لجمهور المستهلكين عامةً في يناير ٢٠١٧. تقتصر التفاصيل الفنية الوحيدة المتاحة لنا حاليًّا على ما ورد في براءة الاختراع التي قُدِّمَت منذ عامَين، والتي تصف نظامًا يُلغي الحاجة إلى التحقُّق من شراء كل عنصر على حدة. وبدلًا من ذلك، تُضاف تفاصيل عربة تسوُّق العميل الحقيقية تلقائيًّا إلى عربة تسوُّقه الافتراضية أثناء التسوق. ويتم الدفع إلكترونيًّا أثناء مغادرة العميل المتجر عبر منطقة انتقالية ما دام يمتلك حسابَ أمازون وهاتفًا ذكيًّا يحتوي على تطبيق أمازون جو. يعتمد نظام جو على مجموعة من أجهزة الاستشعار، عدد كبير جدًّا منها، تُستخدَم لتحديد متى تُؤخَذ سلعة من أحد الرفوف أو تُعاد إليه.

من شأن هذا النظام أن ينتج كميةً هائلة من البيانات ذات الفائدة التجارية لصالح شركة أمازون. وبما أن كلَّ فعل تسوُّق يحدث ما بين دخول العميل المتجر ومغادرته يُسجَّل، فلا شك في أن أمازون ستتمكَّن من استخدام هذه البيانات في تقديم التوصيات لعملائها خلال زيارتهم التالية بطريقة تماثل نظام توصياتها عبر الإنترنت. ولكن، قد تطرأ مشكلات تتعلَّق بمدى تقديرنا لخصوصيتنا، لا سيَّما بسبب أمور على غرار الاحتمالية المذكورة في طلب الحصول على براءة الاختراع، والتي تتعلق باستخدام أنظمة التعرُّف على الوجوه في تحديد العملاء.

نتفليكس

ثمَّة شركة أخرى من شركات وادي السيليكون وهي شركة نتفليكس التي أُسِّسَت عام ١٩٩٧ كشركة لتأجير أقراص الدي في دي عبر البريد. كان بإمكانك أن تأخذ قرص دي في دي وتضيف قرصًا آخر إلى قائمة طلباتك؛ ومن ثمَّ، تُرسل إليك الأقراص تباعًا. والأهم من ذلك أنه في مقدورك تحديد الأولويات ضمن قائمتك. لا تزال هذه الخدمة متوافرةً ومربحة، ولكن يبدو أنها توشك على الانتهاء تدريجيًّا. أصبحت نتفليكس حاليًّا مزوِّدًا دوليًّا للخدمات الإعلامية والبَث عبر الإنترنت، ووصلَ عددُ مشتركيها إلى ما يقارب ٧٥ مليون مشترك من ١٩٠ دولة، وتمكَّنت من التوسُّع بنجاح لتشرع في تقديم برامجها الأصلية.

تجمع نتفليكس كمياتٍ هائلةً من البيانات وتستخدمها في تحسين الخدمات المُقدَّمة إلى العملاء، مثل عرض التوصيات لأفراد المستخدمين مع السعي في الوقت نفسه إلى تقديم خدمة بَث لأفلامها يمكن

التعويل عليها. تقع التوصيات في صميم نموذج عمل شركة نتفليكس، حيث يقوم الجزء الأكبر من عملها على التوصيات التي يمكنها عرضها على العملاء والمستنِدة إلى البيانات. تُتابع نتفليكس حاليًّا ما يشاهده كلُّ عميل من عملائها، وما يتصفحه، وما يبحث عنه، ويومَ وتوقيت أدائه لكل هذه الأنشطة. كما أنها تسجِّل ما إذا كان العميل يستخدم جهاز آي باد، أو تلفزيونًا أو جهازًا آخر.

في ٢٠٠٦، أعلنت نتفليكس عن مسابقة عامة للجمهور تهدف إلى تحسين أنظمة التوصية لديها. وعرضت الشركة جائزةً قيمتها مليون دولار لخوارزمية التصفية التعاونية التي ستُحسِّن بنسبة ١٠ بالمائة من دقة التنبؤات بتقييمات المستخدمين للأفلام. وأتاحت نتفليكس بيانات التدريب، ما يزيد على ١٠٠ مليون عنصر، من أجل مسابقة تعلُّم الآلة والتنقيب في البيانات هذه، ولم يكن مسموحًا استخدام أي مصادر أخرى. عرضت نتفليكس جائزةً مبدئية (جائزة التقدُّم) بقيمة ٥٠ ألف دولار، والتي فازَ بها فريقُ شركة كوربل في ٢٠٠٧ عندما تمكنوا من حَلِّ مشكلةٍ ذات صلة ولكنها أسهل نوعًا ما. كلمة «أسهل» كلمة نسبية في هذا السياق؛ فقد دمجَ حلُّهم ١٠٧ خوارزمياتٍ مختلفة ليحصلوا على خوارزميتَين نهائيتَين لا زالت نتفليكس تستخدمهما حتى الآن، ولا تزالان قيد التطوير المستمر. صيغَت هاتان الخوارزميتان لتتمكَّنا من التعامل مع ١٠٠ مليون تقييم في مقابل الخمسة مليارات تقييم التي يجب أن تتمكَّن الخوارزمية التي ستحصل على الجائزة الكاملة من التعامُل معها. مُنِحَت الجائزة الكاملة، في نهاية المطاف، في عام ٢٠٠٩ إلى فريق برجماتيك كيوس من شركة بلكور، والذي حقَّقت خوارزميته نسبة تحسُّن بلغت ١٠٫٠٦ في المائة عن الخوارزمية التي كانت مستخدَمةً في ذلك الحين. لم تنفِّذ شركة نتفليكس الخوارزمية الفائزة بالكامل على الإطلاق، ويرجع ذلك في الأساس إلى أنها، بحلول ذلك الوقت، كانت قد غيَّرت نموذج عملها إلى نموذج عمل البَث الإعلامي الحالي المألوف.

بمجرد أن وسَّعت نتفليكس نموذج عملها من الخدمات البريدية إلى إتاحة الأفلام عبر البَث، تمكَّنت من جمع كمٍّ أكبر بكثير من المعلومات عن تفضيلات عملائها وعادات المشاهدة، الأمر الذي مكَّنها من تقديم توصياتٍ مُحسَّنة. ولكن، بعيدًا عن الطريقة الرقمية، تُوظِّف نتفليكس مُعلِّقين بدوام جزئي، بإجمالي حوالي ٤٠ شخصًا في جميع أنحاء العالم، يشاهدون الأفلام ويعلِّقون على المحتوى ويصنِّفونه، على سبيل المثال، على أنه «خيال علمي» أو «كوميديا». ومن ثَمَّ تُصنَّف الأفلام، بالاحتكام إلى الرأي البشري في الأساس وليس إلى خوارزمية حاسوبية، وسنتناول هذا لاحقًا.

تستخدم نتفليكس مجموعةً كبيرة من خوارزميات التوصية، والتي تُشكِّل معًا نظام التوصية. تعمل كلُّ هذه الخوارزميات وفقًا للبيانات الضخمة المُجمَّعة التي تجمعها الشركة. على سبيل المثال، تحدِّد التصفية المستندة إلى المحتوى البيانات التي يقدِّمها «المعلقون»، وتبحث عن أفلام وبرامج تلفزيونية مشابهة طبقًا لمعايير على غرار المحتوى أو المُمثِّل. ترصد خوارزميات التصفية التعاونية هذه الأمور على أنها عاداتك فيما يخصُّ المشاهدة والبحث. وتستند التوصيات إلى ما شاهده المشاهِدون أصحاب ملفات التعريف المشابهة. ولا شكَّ في أن فرص نجاح هذا الأسلوب تتراجع عندما يزيد عدد مستخدِمي الحساب عن مستخدم واحد، عادةً ما يكونون عدة أفراد من أسرة واحدة، لكلٍّ منهم أذواقٌ وعاداتٌ مشاهدة مختلفة. ومن أجل التغلُّب على هذه المشكلة، أنشأت نتفليكس خيارَ ملفات التعريف المتعدِّدة ضمن كلِّ حساب من حسابات المستخدِمين.

تَعَد خدمة العروض التلفزيونية على الإنترنت حسب الطلب مجالًا آخر يساعد في نمو شركة نتفليكس، وستتزايد أهمية استخدام أساليب تحليل البيانات الضخمة مع استمرارها في تطوير أنشطتها. بالإضافة إلى جمع بيانات البحث والتقييمات باستخدام النجوم، يمكن لشركة نتفليكس حاليًّا أن تحتفظ بسجلاتٍ عن عدد مرات إيقاف المستخدمين لمقاطع الفيديو أو تقديمها، أو ما إذا كانوا يستكملون مشاهدة كل برنامج بدءوا في مشاهدته أم لا. كما أنها تتابع كيف، ومتى، وأين شاهدوا البرنامج، إلى جانب عدد كبير من المتغيِّرات التي لن يسعنا ذكرها هنا لكثرتها. باستخدام أساليب تحليل البيانات الضخمة، نما إلى علمنا أنها أصبحت قادرةً حاليًّا على التنبؤ بدقة معقولة ما إذا كان أحد عملائها بصدد إلغاء اشتراكه.

عِلمُ البيانات

«عالِم البيانات» هو اللَّقب العام الذي يُطلَق على العاملين في مجال البيانات الضخمة. ألقى تقرير شركة ماكنزي لعام ٢٠١٢ الضوءَ على نقص عدد علماء البيانات في الولايات المتحدة الأمريكية وحدها، مُقدِّرًا أنه بحلول ٢٠١٨ سيصل العجز إلى ١٩٠ ألفًا. يتكرَّر هذا النمط على نحو واضح في جميع أنحاء العالَم، وعلى الرغم من المبادرات الحكومية التي تُشجِّع على التدريب على مهارات عِلم البيانات، يبدو أن الفجوة بين الخبرات المتاحة والمطلوبة لا تزال تتسع. تتزايد شهرة عِلْم البيانات كأحد خيارات الدراسة الجامعية، إلا أن الخريجين لم يتمكَّنوا حتى الآن من تلبية متطلَّبات التجارة والصناعة، حيث تقدِّم الوظائفُ في مجال علم البيانات رواتبَ عاليةً للمتقدمين الأكثر خبرة. تهتم البياناتُ الضخمة للمؤسسات التجارية بالربح، وسرعان ما ستتسلَّل خيبة الأمل إلى نفس محلِّل بياناتٍ مُثقَل بالأعباء ولا يملك الخبرة الكافية إذا فشل في تحقيق النتائج الإيجابية المتوقَّعة. تطلب الشركات، في أغلب الأحيان، نموذجَ عالِم بياناتٍ يفي بكل المتطلَّبات، حيث تُريده ضليعًا في جميع المهام، وتتوقَّع منه أن يتمتَّع بالكفاءة في كل شيء، بدءًا من التحليل الإحصائي وحتى تخزين البيانات وأمن البيانات.

يحظى أمنُ البيانات بأهمية كبيرة بالنسبة إلى أي شركة، وللبيانات الضخمة مشكلاتها الأمنية الخاصة. في ٢٠١٦، أُلغيت مبادرة جائزة نتفليكس الثانية بسبب مخاوف تتعلَّق بأمن البيانات. شملت عمليات اختراق البيانات مؤخرًا شركة أدوبي في ٢٠١٣، وإيباي وبنك جيه بي مورجان تشيس في ٢٠١٤، وشركة أنثيم (شركة تأمين صحي يقع مقرُّها في الولايات المتحدة) وشركة كارفون ويرهاوس في ٢٠١٥، وموقع مَاي-سبيس في ٢٠١٦، وموقع لينكد-إن الذي تعرَّضَ لعملية اختراقٍ وقعت في ٢٠١٢ ولم تُكتشَف حتى ٢٠١٦. وما الشركات السالفة الذكر إلا عينة صغيرة؛ فثمة الكثير من الشركات الأخرى التي تعرَّضت للاختراق أو عانت من أنواع أخرى من الانتهاكات الأمنية التي أدَّت إلى نشر غير مصرَّح به لبياناتٍ حسَّاسة. في الفصل السابع، سنتناول بتعمُّق بعض الانتهاكات الأمنية للبيانات الضخمة.

الفصل السابع

أمن البيانات الضخمة وقضية سنودن

في يوليو ٢٠٠٩، وجدَ قُراء تطبيق أمازون كيندل أن الحياة تحاكي الفَن عندما اختفت نسخُ رواية جورج أورويل «١٩٨٤» تمامًا من أجهزتهم. في رواية ١٩٨٤، يُستخدَم «ثَقب الذاكرة» في حرق المستندات التي تُعَد هدَّامةً أو لم تعد مطلوبة. ومن ثَمَّ، تختفي المستندات إلى الأبدِ وتُعاد كتابة التاريخ. كان من الممكن ألَّا يكون ذلك الأمر سوى مزحة مؤسفة، ولكن، في الحقيقة أزيلت روايتَا «١٩٨٤» و«مزرعة الحيوان» لجورج أورويل من التطبيق بسبب نزاعٍ بين شركة أمازون والناشِر. شعرَ العملاءُ بالغضب لأنهم دفعوا مقابل هذه الكتب الإلكترونية، وافترضوا أنها بذلك أصبحت ملكًا لهم. ورفعَ طالبٌ في المرحلة الثانوية وشخصٌ آخر قضيةً تمَّت تسويتها خارج المحكمة. في هذه التسوية، صرَّحت شركة أمازون بأنها لن تمحو مجدَّدًا كتبًا من تطبيقات كيندل المُثبتة على أجهزة العملاء إلا في ظروف معينة، بما في ذلك وجود «أمر قضائي أو رقابي يتطلَّب هذا الحذف أو التعديل». عرضت أمازون على عملائها استعادة المبالغ المدفوعة، أو الحصول على قسائم هدايا، أو استعادة الكتب المحذوفة. زد على ذلك أننا لا يمكننا بيع الكتب التي اشتريناها على تطبيق كيندل أو إقراضها، ومن ثمَّ يبدو أننا لا نملكها من الأساس.

على الرغم من أن واقعة كيندل كانت بسبب مشكلة قانونية ولم تكن نابعةً عن سوء نية، فإنها تكشف عن مدى سهولة حذف المستندات الإلكترونية، وكيف أنه دون وجود النسخ المطبوعة يمكن بسهولة محو أي نص يُرى على أنه غير مرغوب فيه أو هدَّام محوًا تامًّا. إذا أمسكت بنسخة ورقية من هذا الكتاب وقرأتها، فستدرك يقينًا أنها ستظل على حالها كما هي دون تغيير، ولكن إذا قرأت أيَّ شيء على الويب حاليًّا، فلا يمكنك أن تتيقَّن ممَّا أنها ستظل كما هي في الغد أم لا. لا يمكن أن تتيقَّن من شيءٍ على الويب. وبما أن المستندات الإلكترونية يمكن تعديلها وتحديثها دون معرفة المؤلِّف وعِلْمه، فإنه يمكن التلاعُب بها بسهولة. قد يكون هذا الوضع ضارًّا للغاية في العديد من المواقف المختلفة، مثل احتمالية تلاعب شخص بالسجلات الطبية الإلكترونية. حتى إن التوقيعات الرقمية، المصمَّمة للمصادقة على المستندات الإلكترونية، يمكن اختراقها. من شأن ما سبق أن يبرز بعضًا من المشكلات التي تواجه أنظمة البيانات الضخمة، على غرار ضمان أنها تعمل على النحو المطلوب، وإمكانية إصلاحها في حال تعطُّلها، وأنها مقاومة للتلاعُب، ولا يمكن الوصول إليها إلا بواسطة مَن يملكون التصريح الصحيح.

يدور موضوع النقاش الرئيسي في هذا الفصل حول مسألة تأمين الشبكات والبيانات التي تحتوي عليها. وثمَّة إجراءٌ أساسي يُتخذ لحماية الشبكات من الوصول غير المُصرَّح به وهو تثبيت «جدار حماية»، والذي يعزل الشبكات عن الوصول الخارجي غير المُصرَّح به عبر الإنترنت. حتى وإن كانت الشبكات مُؤمَّنةً ضد الهجمات المباشرة، كالفيروسات وأحصنة طروادة على سبيل المثال، قد تظل البيانات المُخزَّنة فيها عرضةً للخطر، خاصةً إذا لم تكن مُشفَّرة. على سبيل المثال، في أحد هذه الأساليب، وهو التصيُّد الاحتيالي، تُجرَى محاولات لإدخال تعليماتٍ برمجية ضارة، ويكون هذا عادةً عن طريق إرسال رسالة إلكترونية تتضمَّن ملفًّا قابلًا للتنفيذ، أو من خلال طلب بيانات

شخصية أو أمنية مثل كلمات المرور. ولكن، يظل الاختراق الإلكتروني هو المشكلة الرئيسية التي تواجه البيانات الضخمة.

تعرَّض متجر تارجت للبيع بالتجزئة للاختراق في ٢٠١٣، وأدَّى هذا إلى سرقة تفاصيل سجلات ما يُقدَّر بنحو ١١٠ مليون عميل، بما في ذلك تفاصيل بطاقات ائتمان ٤٠ مليون شخص. أفادت التقارير أنه بحلول نهاية نوفمبر كان المتسلِّلون قد نجحوا في إقحام برامجهم الضارة في أغلب أجهزة نقاط البيع الخاصة بمتجر تارجت، وتمكَّنوا من جمع سجلات بطاقات العملاء عن طريق معاملاتٍ في الوقت الحقيقي. في ذلك الحين، كان نظام تارجت الأمني يُراقَب على مدار الساعة من قِبَل فريق من المختصين يعمل في بنجالور. جرى التنبيه بوجود نشاط مشبوه وتواصَل الفريقُ مع فريق الأمنِ الرئيسي في مينيابوليس، الذي لم يتخذ، للأسف، أيَّ إجراءاتٍ بمقتضى هذه المعلومات. كان اختراقُ متجر هوم ديبوت، الذي سنتناوله في الفقرة التالية، أكبر بكثير، ولكنه استخدمَ أساليبَ مشابهةً أدَّت إلى سرقة كمية هائلة من البيانات.

اختراق متجر هوم ديبوت

في ٨ سبتمبر ٢٠١٤، أعلن متجر هوم ديبوت، الذي يصف نفسه بأنه أكبر متجر بيع بالتجزئة لمستلزمات تحسين المنازل في العالم، في بيان صحفي، أن نظم بيانات الدفع قد تعرَّضت للاختراق. وفي استكمال للبيان في ١٨ سبتمبر ٢٠١٤، أعلنَ هوم ديبوت أن الهجوم قد أثَّر على حوالي ٥٦ مليون بطاقة خصم مباشر وائتمان. بعبارة أخرى، سُرقت تفاصيل ٥٦ مليون بطاقة خصم مباشر وائتمان. علاوةً على ما سبق، سُرقَت عناوين ٥٣ مليون موقع إلكتروني. في هذه الحالة، تمكَّن المخترقون من سرقة سجل أحد الموردِّين أولًا، ما أتاح لهم إمكانية الوصول بسهولة إلى النظام، ولكن، إلى جزء النظام المتعلِّق بهذا المورِّد فقط. وأُجريت عملية الاختراق هذه عن طريق محاولة تصيُّد احتيالي ناجحة.

استلزمت الخطوة التالية أن يتمكَّن المخترقون من الوصول إلى النظام بالكامل. وفي هذه المرة، كان هوم ديبوت يستخدم نظام تشغيل مايكروسوفت إكس بي، والذي كان يحتوي على خطأٍ جوهري استغله المخترقون. استُهدف بعد ذلك نظامُ الدفع الذاتي؛ لأن هذا النظام الفرعي كان يمكن تحديده بوضوح ضمن النظام ككل. وأخيرًا، أصابَ المخترقون أجهزة الدفع الذاتي البالغ عددُها ٧٥٠٠ جهاز ببرنامج ضار ليحصلوا على معلومات العملاء. استخدم المخترقون BlackPOS، الذي يُعرف أيضًا باسم «كابتوكسا»، وهو برنامج ضار مُختص في استخراج معلومات بطاقات الخصم المباشر والائتمان من الوحدات الطرفية المتضرِّرة. لأغراض الأمان، يجب أن تُشفَّر معلومات بطاقة الدفع عند تمريرها على إحدى المحطات الطرفية لنقاط البيع، ولكن، يبدو أن هذه الخاصية المعروفة باسم التشفير من نقطة إلى نقطة لم تكن مُفعَّلة؛ ومن ثمَّ، أصبحت التفاصيل متاحةً أمام المخترقين ليستَولوا عليها.

اكتَشفَت هذه السرقة عندما بدأت البنوك تكتشف أنشطةَ احتيالية لحساباتٍ كانت قد أجرت عمليات شراء أخرى من متجر هوم ديبوت منذ فترة قصيرة، كانت قد بيعت تفاصيل البطاقات عبر ريسكاتور، منفذ جرائم إلكترونية موجود على الويب المظلم (دارك ويب). المثير في الأمر أن الأشخاص الذين استخدموا آلات تسجيل النقد، التي تتطلَّب أيضًا استخدام البطاقات، لم يتأثروا بهذا الهجوم. ويبدو أن السبب في ذلك أن آلات تسجيل النقد يتعرَّف عليها الكمبيوتر المركزي عن طريق الأرقام فقط، ولا يمكن للمجرمين التعرُّف عليها بسهولة بوصفها نقاط دفع. لو حدثَ أنَّ هوم ديبوت استخدمَ أيضًا الأرقام البسيطة مع وحدات الدفع الذاتي الطرفية لديه، فلربما نجحَ في إحباط محاولة الاختراق هذه. وعلى ذكر هذا، فقد كان نظام كابتوسكا في ذلك الوقت واحدًا من البرامج الضارة المتطوِّرة ولا يمكن اكتشافه تقريبًا؛ ومن ثمَّ فإنه في ضوء الوصول المفتوح إلى النظام الذي أتاحها للمخترقين، تمَّ إدخاله إلى النظام بنجاح في نهاية المطاف.

أكبر اختراق للبيانات على الإطلاق

في ديسمبر ٢٠١٦، أعلنت شركة ياهو أن اختراقًا للبيانات يتضمَّن ما يزيد على المليار مستخدم وقعَ في أغسطس ٢٠١٣. في هذا الاختراق الذي أُطلق عليه أكبر عملية سرقة إلكترونية للبيانات الشخصية على الإطلاق، أو على الأقل أكبر عمليةٍ أعلنت عنها أي شركة على الإطلاق، يبدو أن اللصوص استخدموا ملفات تعريف ارتباط زائفة مكَّنتهم من الوصول إلى الحسابات من دون الحاجة إلى كلمات مرور. جاءَ هذا الاختراق بعد الإفصاح عن هجوم وقعَ على ياهو في ٢٠١٤ اختُرقت خلاله حسابات ٥٠٠ مليون مستخدِم. المفاجئ في الأمر أنَّ ياهو زعمت أن الاختراق الذي حدث في ٢٠١٤ قد دبَّرته «جهة ترعاها الدولة» لم تُفصح عن اسمها.

أمن الحَوْسَبة السحابية

تزداد قائمة الاختراقات الأمنية للبيانات الضخمة كلَّ يوم تقريبًا. وأضحت سرقة البيانات، واحتجاز البيانات مقابل طلب فِدية، وتخريب البيانات، مخاوف كبرى في عالَمنا الحالي القائم في أساسه على البيانات. ثمَّة الكثير من المخاوف المتعلِّقة بأمن البيانات الشخصية الرقمية وملكيتها. قبل العصر الرقمي كنا نحتفظ بالصور في ألبومات، وكان نيجاتيف الصور هو نسختنا الاحتياطية. بعد ذلك، أصبحنا نخزِّن صورنا إلكترونيًّا على الأقراص الصلبة لأجهزة الكمبيوتر الخاصة. ولأن أجهزة الكمبيوتر كانت عُرضةً لأن تتعطَّل، اقتضت الحكمة أن نحتفظ بنسخ احتياطية، ولكن، على الأقل لم تكن الملفات متاحةً للجميع. أصبح الكثير منا الآن يُخزِّنون البيانات في السحابة الإلكترونية. ونظرًا لِمَا تتطلَّبه الصور، ومقاطع الفيديو، والأفلام المنزلية من مساحة تخزين كبيرة، فإن السحابة الإلكترونية بدت منطقيةً من هذا المنظور. عندما تخزِّن ملفاتِك في السحابة الإلكترونية، فإنك ترفعها

إلى مركز بيانات — بل إنها تَوزّع، على الأرجح، على عدة مراكز بيانات — ومن ثمَّ يُحتفَظ بأكثر من نسخة واحدة منها.

إذا خزَّنت كلَّ صورك في السحابة الإلكترونية، فمن غير الوارد على الإطلاق، بفضل الأنظمة المتطوِّرة المعاصرة، أن تفقدها. وعلى النقيض، إذا أردت أن تحذف شيئًا، ربما صورةً أو مقطع فيديو، فمن الصعب أن تتأكَّد من أن كل النسخ قد حُذِفَت. وسيكون عليك أن تعتمد بصفة أساسية على مزوِّد الخدمة في ذلك. ثمَّة موضوع مهم آخر وهو التحكم فيمَن يُتاح لهم الوصول إلى هذه الصور وغيرها من البيانات التي رفعتها إلى السحابة الإلكترونية. إذا أردنا تأمين البيانات الضخمة، فلا بد من التشفير.

التشفير

يشير التشفير، كما ذكرنا باختصار في الفصل الخامس، إلى الأساليب المُستخدَمة في خلط الملفات حتى لا يمكن قراءتها بسهولة، ويعود الأسلوب الأساسي إلى العصر الروماني على أقل تقدير. يصف جايوس سويتونيوس، في كتابه «القياصرة الاثنا عشر»، كيف أن يوليوس قيصر شفَّر الوثائق عن طريق إزاحة الحروف بمقدار ثلاثة حروف إلى اليسار. باستخدام هذا الأسلوب، تُشفَّر كلمة secret إلى pbzobq. تُعرَف هذه الشفرة باسم «شفرة القيصر»، وهي شفرة ليس من الصعب فكُّها، إلا أنَّ أكثر الشفرات المُستخدَمة حاليًّا أمانًا تُطبِّق أسلوب الإزاحة كجزء من الخوارزمية المُستخدَمة.

في ١٩٩٧، أُثِبتَ أفضل أسلوب تشفير متاح للعامة، وهو معيار تشفير البيانات (دي إي إس)، أنَّ من الممكن فكَّ شفرته، ويرجع هذا بدرجة كبيرة إلى زيادة القدرة الحاسوبية المتاحة وطول مفتاح التشفير القصير نسبيًّا الذي يبلغ ٥٦بت. على الرغم من أن هذا الأسلوب يتيح $٢^{٥٦}$ من اختيارات المفاتيح المختلفة الممكنة، فإنه يمكن فكُّ شفرة الرسائل عن طريق اختبار كل مفتاح إلى حين العثور على المفتاح الصحيح. وهذا ما حدث بالفعل عام ١٩٩٨، في أقل من ٢٤ ساعةً باستخدام جهاز ديب كراك، وهو جهاز كمبيوتر صمَّمته مؤسسة الحدود الإلكترونية خصوصًا لهذا الغرض.

في ١٩٩٧، أجرى المعهد الوطني للمعايير والتقنية في الولايات المتحدة الأمريكية، لتخوُّفه من أن معيار تشفير البيانات يفتقر إلى الأمان اللازم لحماية الوثائق الفائقة السرية، مسابقةً مفتوحة على مستوى العالَم للتوصل إلى أسلوب تشفير أفضل من معيار تشفير البيانات. انتهت المسابقة في ٢٠٠١ باختيار خوارزمية معيار التشفير المتقدِّم. قُدِّمَت الخوارزمية تحت مُسمًّى خوارزمية ريندايل، الذي دمجَ بين اسمَي مُبتكرَيها البلجيكيَّين جون دايمن وفينسنت ريمن.

معيار التشفير المتقدِّم عبارة عن خوارزمية برمجية لتشفير النصوص يمكن من خلالها الاختيار من بين مجموعة من مفاتيح التشفير الأكثر طولًا: ١٢٨بت، أو ١٩٢بت، أو ٢٥٦بت. بالنسبة إلى طول مفتاح التشفير البالغ ١٢٨بت، تحتاج الخوارزمية إلى تسع جولات معالجة تتكوَّن كل منها من أربع

خطوات، بالإضافة إلى جولة أخيرة مكَونة من ثلاث خطوات فقط. يجري تنفيذ خوارزمية معيار التشفير المتقدِّم على نحو تكراري، وتُجري عددًا كبيرًا من العمليات الحاسوبية على مصفوفات، فقط نوع العمليات الحسابية الذي من الأفضلَ إجراؤه باستخدام أجهزة الكمبيوتر. ولكن، يمكننا أن نصف العملية على نحو غير متخصِّص من دون التطرُّق إلى ذلك التحويلات الرياضية.

يبدأ معيار التشفير المتقدِّم بتطبيق مفتاح تشفير على النص الذي نرغب في تشفيره. بعد ذلك لن نتمكَّن من تمييز النص، ولكن بما أننا نعرف مفتاح التشفير، يمكننا أن نفك تشفير النص بسهولة؛ ومن ثمَّ يستلزم الأمر مزيدًا من الخطوات. تتضمَّن الخطوة التالية استبدال كل حرف بحرفٍ آخر باستخدام جدول مرجعي خاص يُسمَّى مربع ريندايل للاستبدال. ومجدَّدًا، إذا كان لدينا مربع ريندايل للاستبدال، فيمكننا العمل على نحو عكسي وفكُّ تشفير الرسالة. تُشكِّل شفرة القيصر، التي تتم فيها إزاحة الحروف إلى اليسار، وعملية استبدال أخيرة للحروف إحدى الجولات. يُستخدَم الناتج بعد ذلك في بدء جولة أخرى، باستخدام مفتاح مختلف وهكذا، حتى تكتمل جميع الجولات. وبالطبع يجب أن نكون قادرين على فك الشفرة، وفيما يخصُّ هذه الخوارزمية يمكن أن تعكس هذه العملية.

بالنسبة إلى مفتاح التشفير البالغ طوله ١٩٢بت، ثمة ١٢ جولةً إجمالًا. ولمزيدٍ من الأمان، وهو ما يتحقَّق باستخدام مفتاح تشفير أطول، يمكن استخدام مفتاح التشفير البالغ طوله ٢٥٦بت، إلا أن أغلب المستخدمين، بما في ذلك جوجل وأمازون، يرَون أن مفتاح التشفير الذي طوله ١٢٨ كافٍ لتلبية المتطلبات الأمنية لبياناتهم الضخمة. إن معيار التشفير المتقدِّم آمن، ولم يتمكَّن أحدٌ من اختراقه حتى الآن، ما جعل العديد من الحكومات تطلب من شركاتٍ كبرى — مثل أبل وجوجل — أن تتيح مداخل سريةً إلى المادة المُشفَّرة.

أمن البريد الإلكتروني

تشير التقديرات إلى أنه في ٢٠١٥، كان يُرسَل ما يزيد على ٢٠٠ مليار رسالة إلكترونية كلَّ يوم، وكانت نسبة تقل عن ١٠ بالمائة منها فقط موثوقةً وليست بريدًا عشوائيًّا أو ذات نوايا خبيثة. وتكون أغلب الرسائل الإلكترونية غير مشفَّرة، ما يجعل محتواها عُرضةً لأن يترصَّده المخترقون. عندما أرسلُ رسالةً إلكترونية غير مشفَّرة، من كاليفورنيا إلى المملكة المتحدة على سبيل المثال، فإنها تُقسَّم إلى «حزم» من البيانات وتُنقَل عبر خادم بريدٍ متصل بالإنترنت. يتكوَّن الإنترنت في الأساس من شبكة عالمية ضخمة من الأسلاك الموجودة فوق الأرض، وتحت الأرض، وتحت المحيطات، بالإضافة إلى أبراج الهواتف المحمولة والأقمار الصناعية. والقارة الوحيدة غير المُوصَّلة بكابلاتٍ عابرة للمحيطات هي القارة القطبية الجنوبية (أنتاركتيكا).

ومن ثمَّ، على الرغم من الاعتقاد الشائع بأن الإنترنت والحَوْسَبة المستندة إلى السحابة الإلكترونية لا سلكيان، فإنها ليست كذلك على الإطلاق؛ فالبيانات تُنقَل عبر كابلات ألياف ضوئية ممدودة تحت المحيطات. وتُنقَل جميع الاتصالات الرقمية بين القارات تقريبًا بهذه الطريقة. ستُرسَل رسالتي الإلكترونية عبر كابلات ألياف ضوئية عابرة للمحيط الأطلنطي، حتى وإن كنت أستخدم خدمة

حوسبة سحابية. تستحضر السحابة الإلكترونية، تلك الكلمة الجذابة الواسعة الانتشار، إلى الذهن، صورًا لأقمارٍ صناعيةٍ ترسل البيانات إلى جميع أنحاء العالم، ولكن، في الواقع تمتد جذور الخدمات السحابية عميقًا في شبكة مُوزَّعة من مراكز البيانات التي توفِّر الوصول إلى الإنترنت، عبر كابلاتٍ في الغالب.

توفِّر كابلات الألياف الضوئية أسرع وسيلة لنقل البيانات، وعليه، فهي تحظى بأفضلية على الأقمار الصناعية بوجه عام. وينتج عن الأبحاث المُكثَّفة في تقنية الألياف الضوئية سرعات نقل بيانات أعلى من أيِّ وقتٍ مضى. لطالما كانت الكابلات العابرة للمحيط الأطلنطي هدفًا لهجمات غريبة وغير متوقَّعة، بما في ذلك هجمات أسماك القرش بقصد قضمها وقطعها. على الرغم من أن هجمات أسماك القرش على الكابلات، طبقًا للجنة الدولية لحماية الكابلات، مسئولة فقط عن أقل من ١٪ من الأعطال المُسجَّلة، أصبحت الكابلات في المناطق الأكثر عُرضةً للخطر تُحمَى حاليًّا باستخدام ألياف الكيفلار. بافتراض أن الكابلات العابرة للمحيط الأطلنطي لا تواجه أيَّ مشكلاتٍ مع أسماك القرش الفضولية، أو الحكومات المُعادية، أو صيادي الأسماك غير المكترثين، وأن رسالتي الإلكترونية بلغت البر الرئيسي للمملكة المتحدة وتواصل طريقها، ربما تتعرَّض في هذه المرحلة، على غرار غيرها من بيانات الإنترنت، للاعتراض. في يونيو ٢٠١٣، سرَّبَ إدوارد سنودن مستنداتٍ تكشف عن أن مكاتب الاتصالات الحكومية في المملكة المتحدة تتنصَّت على كمياتٍ هائلة من البيانات التي تصل البلاد عبر حوالي ٢٠٠ كابل عابر للمحيط الأطلنطي، باستخدام نظام يُسمَّى تِمبورا.

قضية سنودن

إدوارد سنودن خبيرٌ أمريكيٌّ محترف في أجهزة الكمبيوتر، اتُّهم بالتجسُّس في ٢٠١٣ بعد أن سرَّب معلوماتٍ سريةً من وكالة الأمن القومي الأمريكية. وضعت هذه القضية الذائعة الشهرة إمكاناتِ المراقبة الشاملة للحكومة تحت منظار عامة الشعب، وأُعربَ على نطاق واسع عن مخاوف تتعلَّق بخصوصية الأفراد. حصلَ سنودن على الكثير من الجوائز منذ أن أقدم على هذا الفعل، والتي شملت انتخابه رئيسًا لجامعة جلاسكو، وجائزة شخصية العام من جريدة «الجارديان» لعام ٢٠١٣، والترشُّح لجائزة نوبل للسلام عن الأعوام ٢٠١٤ و٢٠١٥ و٢٠١٦. كما حصل على دعم منظمة العفو الدولية بوصفه شخصًا قدَّم خدمةً لبلده بوصفه أحد كاشِفي الفساد. ولكن، يعارض المسئولون الحكوميون والسياسيون الأمريكيون هذا الرأي.

في يونيو ٢٠١٣، أفادت جريدة «الجارديان» في المملكة المتحدة بأن وكالة الأمن القومي الأمريكية تجمع بيانات تعريفٍ من عددٍ من شبكات الهواتف الكبرى في الولايات المتحدة. وسرعان ما أعقبَ هذا التقرير الكشفُ عن برنامج يُسمَّى بريزم، والذي كان يُستخدم في جمع بيانات من الإنترنت وتخزينها، تتعلَّق بمواطنين أجانب يتواصلون مع أشخاص داخل الولايات المتحدة. بعد ذلك، ظهر عدد كبير جدًّا من التسريبات التي تُدين كلًّا من الولايات المتحدة والمملكة المتحدة. كان إدوارد سنودن، موظف شركة بوز ألين هاميلتون ومتعاقدًا مع وكالة الأمن القومي الأمريكية الذي يعمل في مركز هاواي للتشفير، هو مصدر هذه التسريبات التي أرسلها إلى إعلاميين وظنَّ أنه يمكنه الوثوق

في أنهم لن ينشروها دون دراسة متأنية. لا يتسع المجال في هذا الكتاب لذكر دوافع سنودن والمسائل القانونية المتضمَّنة، ولكن من الواضح أنه كان يعتقد أن ما بدأ كتجسُّس مشروع على الدول الأخرى قد انقلبَ على نفسه، وأصبحت وكالة الأمن القومي الأمريكية تتجسَّس، بطريقة غير قانونية، على جميع المواطنين الأمريكيين.

توفِّر أداتا تجريف الويب، DownThemAll التي هي ملحق متاح لمتصفح موزيلا فايرفوكس، وبرنامج wget، وسيلةً للتنزيل السريع لكامل محتويات المواقع الإلكترونية أو غيرها من بيانات الويب. استخدم سنودن هذَين التطبيقَين، المتاحَين للمستخدمين المُصَّرَح لهم بالوصول إلى شبكاتٍ وكالة الأمن القومي السرية، في تنزيل كمياتٍ هائلة من المعلومات ونسخها. كما نقلَ كمياتٍ ضخمةً من البيانات الشديدة الحساسية من نظام كمبيوتر إلى آخر. ولكي يتمكَّن من القيام بذلك، كان بحاجة إلى أسماء المستخدمين وكلمات المرور التي يحتفظ بها مديرو الأنظمة عادة. ومن ثمَّ، تمكَّن من الوصول بسهولة إلى الكثير من المستندات السرية التي سرقها، ولكن ليس جميعها. ولكي يحصل على المستندات الفائقة السرية، كان عليه استخدام تفاصيل المصادقة الخاصة بحسابات المستخدمين ذات المستوى الأعلى، الأمر الذي كان من المفترض أن تحول بروتوكولات الأمان دون حدوثه. ولكن، بما أنه مَن أنشأ هذه الحسابات ويمتلك امتيازات مسئول النظام، كان يعلم تفاصيل هذه الحسابات. تمكَّن سنودن أيضًا من إقناع موظف واحد على الأقل من موظفي وكالة الأمن القومي، ممَّن يملكون تصريحاتٍ أمنيةً أعلى ممَّا يملكها، بأن يخبروه بكلمات مرورهم.

وأخيرًا، نسخَ سنودن حوالي ١٫٥ مليون مستند فائق السرية سُلِّم حواليٍ ٢٠٠ ألف مستندٍ منها (كان سنودن يدرك أنه لا يجدر به نشر جميع المستندات المسروقة علانيةً وكانَ حذرًا للغاية فيما يتعلَّق بالمستندات التي يجب نشرها) إلى مراسلين صحفيين موثوقين، ولكن لم يُنشَر من هذه المستندات إلا عدد قليل نسبيًّا في نهاية المطاف.

على الرغم من أن سنودن لم يُفصح أبدًا عن كامل التفاصيل، فيبدو أنه تمكَّن من نسخ البيانات على محركات أقراص محمولة، لم يواجه صعوبةً في أخذها معه عند مغادرته العمل كل يوم. ومن الجليِّ أن الإجراءات الأمنية التي كان من شأنها أن تمنع سنودن من نقل هذه المستندات لم تكن كافية. كان من شأن التفتيش الجسدي البسيط عند الخروج من المُنشأة أن يكشف أيَّ أجهزة محمولة، كما أن كاميرات المراقبة في المكاتب كانت ستشير إلى وجود نشاط مشبوه. في ديسمبر ٢٠١٦، رفعَ مجلس النواب الأمريكي النقاب عن مستندٍ بتاريخ سبتمبر ٢٠١٦، وكان جزءٌ كبير جدًّا منه محجوبًا، يُقيِّم سنودن كشخص ويُقيِّم كذلك طبيعة المستندات المُسرَّبة وتأثيرها. يتضح من هذا المستند أنَّ وكالة الأمن القومي لم تطبِّق إجراءاتٍ أمنيةً كافية، ونتيجةً لهذا بدأ تطبيق مبادرة تأمين الإنترنت منذ ذلك الحين، ولكنها لم تدخل حيز التنفيذ الكامل.

كان سنودن يمتلك امتيازات مسئول نظام واسعة، ولكن طبقًا للطبيعة الشديدة الحساسية للبيانات، كان السماح لشخص واحد بامتلاك حق الوصول الكامل إليها من دون وجود أي احتياطاتٍ أمرًا غير مقبول. على سبيل المثال، ربما كان طلبُ استيفاءِ بيانات اعتماد شخصَين عند محاولة الوصول إلى البيانات أو نقلها كافيًا لمنع سنودن من نسخ الملفات بطريقة غير مشروعة. ومن الغريب أيضًا أن سنودن تمكَّن من توصيل محرك أقراص «يو إس بي» (الناقل التسلسلي العام) ونسخ أي شيءٍ يريد.

وكان من بين إجراءات الأمان البسيطة للغاية تعطيل منافذ «دي في دي» (أقراص الفيديو الرقمية) و«يو إس بي» (الناقل التسلسلي العام) أو عدم تركيبها من الأساس. كان من شأن إضافة مصادقة إضافية باستخدام مسح شبكية العين إلى طلب كلمة المرور أن يُصَعِّب إلى حَدٍّ كبير على سنودن أن يتمكَّن حتى من الوصول إلى هذه المستندات الفائقة السرية. تتسم أساليب الأمان المعاصرة بأنها متطوِّرة ويصعب اختراقها في حال استخدامها على النحو الصحيح.

في أواخر ٢٠١٦، كان البحث بـ «إدوارد سنودن» على محرك بحث جوجل يعطي أكثر من ٢٧ مليون نتيجة بحث خلال ما يزيد قليلًا عن ثانية واحدة، وكان مصطلح البحث «سنودن» يعطي ٤٥ مليون نتيجة بحث. وبما أن الكثير من هذه المواقع تمنح إمكانية الوصول إلى هذه المستندات المُسرَّبة المُصنَّفة على أنها «سرية للغاية» أو تعرضها، فقد أصبحت بالتأكيد في المجال العام العالمي، ولا شكَّ في أنها ستظل كذلك. ويعيش إدوارد سنودن حاليًّا في روسيا.

على النقيض من قضية إدوارد سنودن، تُمثِّل ويكيليكس قصةً مختلفة تمامًا.

ويكيليكس

ويكيليكس هي منظمةٌ ضخمة لكشف الفساد والإبلاغ عن المخالفات عبر الإنترنت، تهدف إلى نشر المستندات السرية. تُموَّل المنظمة بالتبرعات، وأغلب العاملين بها من المتطوِّعين، ولكن يبدو أنها توظِّف عددًا محدودًا من الموظفين أيضًا. حتى ديسمبر ٢٠١٥، زعمت ويكيليكس أنها نشرت (أو سرَّبت) أكثر من ١٠ ملايين مستند. تحافظ ويكيليكس على صورتها العامة الجيدة عبر موقعها ومن خلال موقعَي تويتر وفيسبوك.

تصدَّرت منظمةُ ويكيليكس، المثيرة للجدل، ورئيسُها جوليان أسانج عناوينَ الصحف في ٢٢ أكتوبر ٢٠١٠ عندما نشرت كمية كبيرة للغاية من البيانات السرية، ٣٩١٨٣٢ مستندًا، تحت عنوان «سِجلات حرب العراق». جاءت هذه المستندات بعد مستنداتٍ يبلغ عددها حوالي ٧٥ ألف مستند تتألَّف منها «يوميات الحرب الأفغانية» التي تسرَّبت بالفعل في ٢٥ يوليو ٢٠١٠.

كان أحد جنود الجيش الأمريكي، برادلي مانينج، هو المسئول عن كلا التسريبَين. كان الجندي يعمل محلِّلًا استخباراتيًّا في العراق، وأخذ معه قرصًا مدمجًا إلى العمل، ونسخ مستنداتٍ سريةً من جهاز كمبيوتر شخصي من المفترض أنه آمن. بسبب هذا الفعل، حُكِم على برادلي مانينج، الذي يُعرف الآن باسم تشيلسي مانينج (بعد تحوُّله جنسيًّا)، في ٢٠١٣ بالسجن لمدة ٣٥ عامًا بعد إدانته من قِبَل المحكمة العسكرية لانتهاكه قانون التجسُّس وجرائم أخرى ذات صلة. وخفَّفَ الرئيسُ الأمريكي السابق باراك أوباما الحُكم على تشيلسي مانينج في يناير ٢٠١٧، قبل ترك منصبه. وأُطلِقَ سراح الآنسة مانينج، التي كانت تُعالج من اضطراب الهُوية الجنسية أثناء فترة حبسها، في ١٧ مايو ٢٠١٧.

على الرغم من الانتقادات الشديدة التي تعرَّضت لها منظمة ويكيليكس من السياسيين والحكومات، فقد أُشيدَ بها وحصلت على جوائز من مؤسسات على غرار منظمة العفو الدولية عام ٢٠٠٩ وجريدة «ذي إيكونوميست» عام ٢٠٠٨، ضمن قائمةٍ مطوَّلة من المنظمات الأخرى. طبقًا لموقع ويكيليكس، رُشِّحَ جوليان أسانج لجائزة نوبل للسلام لستة أعوام متتالية، من ٢٠١٠ إلى ٢٠١٥. لا تُفصح لجنة جائزة نوبل عن أسماء المُرشَّحين لنيلها إلا بعد مرور ٥٠ عامًا، أمَّا أعضاء لجنة الترشيح، الذين يتعيَّن عليهم استيفاءُ المعايير الصارمة للجنة جائزة السلام، فغالبًا ما يُفصحون عن أسماء مُرَشَّحِيهم علنًا. على سبيل المثال، في ٢٠١١، رُشِّح جوليان أسانج من قِبَل البرلماني النرويجي سنور فالن دعمًا منه لمنظمة ويكيليكس على كشفها للانتهاكات المزعومة لحقوق الإنسان. وفي ٢٠١٥، حصلَ أسانج على دعم عضو مجلس النواب البريطاني السابق جورج جالوي، وفي أوائل ٢٠١٦ نادى فريق دعم من الأكاديميين بحصول أسانج على الجائزة.

ولكن، بحلول نهاية ٢٠١٦، تحوَّلت الآراء ضد أسانج وويكيليكس، وهو ما يُعزى جزئيًّا على أقل تقديرٍ إلى مزاعم التحيُّز في تقاريرها. استندت الاعتراضات المُثارة ضد ويكيليكس إلى مخاوف متعلِّقة بأمن الأفراد وخصوصيتهم، وخصوصية المؤسسات، والسرية الحكومية، وحماية المصادر المحلية في مناطق النزاعات، والمصلحة العامة بوجه عام. ثم ازدادت الأوضاع تعقيدًا بالنسبة إلى جوليان أسانج وويكيليكس. على سبيل المثال، في ٢٠١٦، سُرِّبَت رسائلُ إلكترونية في أنسب وقتٍ للإضرار بترشح هيلاري كلينتون للرئاسة، الأمر الذي أثار تساؤلاتٍ تتعلَّق بموضوعية ويكيليكس، وأثار انتقاداتٍ كبيرةً من عدد من المصادر التي تحظى باحترام كبير.

بغض النظر عمَّا إذا كنتَ من المؤيِّدين لأفعال جوليان أسانج وويكيليكس أو المعارضين لها، ولا شكَّ أن هذا هو حال الناس عمومًا حيث تتباين آراؤهم تجاه القضية المطروحة، فإن أحد أهم الأسئلة الفنية المهمة هو ما إذا كان من الممكن حقًّا إغلاق موقع ويكيليكس أم لا. بما أن ويكيليكس تحتفظ ببياناتها على العديد من الخوادم في جميع أنحاء العالم، بعضها في بلدان متعاطفة معها، فمن غير المرجَّح أن يُغلَق الموقع بالكامل، حتى وإن افترضنا أن وجوده غير مرغوب فيه. ولكن، إمعانًا في الحماية من الهجمات الانتقامية بعد كل تسريب، أصدرت ويكيليكس ملف تأمين. يتمثَّل الهدف غير المُفصح عنه لهذا الملف في أنه في حال حدوث أي شيءٍ لأسانج أو إغلاق موقع ويكيليكس، فسوف يُرسَل مفتاح التشفير الخاص بملف التأمين ليصبح متاحًا على الملأ. يستخدم أحدثُ ملف تأمين من ويكيليكس معيارَ التشفير المتقدِّم بمفتاح تشفير ٢٥٦بت؛ ومن ثمَّ فمن غير المرجَّح بدرجة كبيرة أن يتعرَّض للاختراق.

شبَّ خلافٌ بين إدوارد سنودن وويكيليكس منذ عام ٢٠١٦. ويتعلَّق الأمر بالطريقة التي اتبعها كلٌّ منهما في إدارة تسريبات البيانات. كان سنودن قد سلَّمَ ملفاته إلى صحفيين موثوقين انتقوا بتأنٍّ المستندات التي يجب تسريبها. كما أُبلِغَ مسئولون حكوميون أمريكيون بالأمر مقدَّمًا، وبناءً على نصائحهم، لم يُسرَّب المزيد من المستندات بسبب مخاوف تتعلَّق بالأمن القومي. وحتى يومنا هذا، ثمة الكثير من المستندات التي لم يُفصَح عنها. ولكن، يبدو أن ويكيليكس تنشر بياناتها من دون أن تبذل جهدًا كبيرًا لحماية المعلومات الشخصية. ولا تزال ويكيليكس تسعى إلى جمع المعلومات من كاشفي الفساد، ولكن، لم تعد موثوقية تسريبات البيانات الأخيرة واضحة، أو ما إذا كان اختيار المعلومات التي تقدِّمها تشير إلى أنها نزيهة بالكامل. تنشر ويكيليكس، على موقعها، تعليماتٍ تتعلَّق

بكيفية استخدام آلية تَسمَّى تَور (مُوجِّه الطبقات، أو حرفيًّا «الموجِّه البَصَلي») في إرسال البيانات دون الكشف عن الهُوية وضمان الخصوصية، ولكن، لا يشترط بالضرورة أن تكون كاشفَ فسادٍ لكي تستخدم هذه الآلية.

متصفح تُور والويب المظلم

قرَّرت جانيت فيرتيسي، وهي أستاذٌ مساعِد في قسم علم الاجتماع بجامعة برينستون، أن تُجري تجربةً شخصية لترى ما إذا كانت ستتمكَّن من إبقاء مسألة حملها سِرًّا عن المسوِّقين عبر الإنترنت؛ ومن ثمَّ منع أن تصبح معلوماتها الشخصية جزءًا من البيانات الضخمة. في مقالٍ نُشِر في مجلةٍ «تايم» في مايو ٢٠١٤، قصَّت د. فيرتيسي تجربتها. كانت قد اتخذت معايير خصوصية استثنائيةً شملت تجنُّب شبكات التواصل الاجتماعي، ونزَّلت متصفِّح تُور واستخدمته في طلب الكثير من أغراض الأطفال، ودفعت مقابل مشترياتها من المتاجر نقدًا. كان كلُّ ما فعلته قانونيًّا تمامًا، ولكنها استنتجت في نهاية المطاف أن اختيار عدم المشاركة أمرٌ مكلِّف ويستهلك الكثير من الوقت، وجعلها تبدو، طبقًا لكلماتها، «مواطنة سيئة». ولكن، يستحق متصفِّح تُور أن نتناوله بالبحث وبالدراسة، خاصةً أنه جعل الدكتورة فيرتيسي تشعر بالأمان وتُحافظ على خصوصيتها من برامج التتبُّع.

متصفِّح تُور عبارة عن شبكة مُشفَّرة من الخوادم أنشأتها البحرية الأمريكية في الأساس من أجل توفير طريقة لاستخدام الإنترنت دون الكشف عن الهُوية؛ ومن ثمَّ تجنُّب التتبُّع وجمع البيانات الشخصية. ومتصفِّح تُور مشروعٌ مستمر يهدف إلى تطوير وتحسين بيئات إخفاء الهُويات عبر الإنترنت المفتوحة المصدر، والتي يمكن لأيٍّ من المهتمين بالخصوصية استخدامها. يعمل البرنامج عن طريق تشفير بياناتك، بما في ذلك عنوان الإرسال، ثم يُجهِّلها عبر إزالة جزء من العنوان، بما في ذلك عنوان بروتوكول الإنترنت بالأساس؛ لأن الشخص يمكن أن يُعثر عليه بسهولة عن طريق التتبُّع العكسي بِناءً على هذه المعلومات. بعد ذلك، تُوجَّه حُزمة البيانات الناتجة عبر نظام من الخوادم أو المُرحِّلات، التي يستضيفها متطوِّعون، قبل أن تصل إلى وجهتها الأخيرة.

تتمثَّل أوجه الاستخدام الإيجابية لمتصفِّح تُور في استخداماته من قِبَل قوات البحرية الأمريكية الذين صمَّموه في الأساس، وصحفيي التحقيقات الذين يرغبون في حماية مصادرهم ومعلوماتهم، والمواطنين العاديين الذين يرغبون في حماية خصوصيتهم. تستخدم الشركات متصفِّح تُور من أجل الاحتفاظ بالأسرار التجارية وإخفائها عن الشركات الأخرى، وتستخدمه الحكومات في حماية مصادر المعلومات الحسَّاسة بالإضافة إلى المعلومات نفسها. قدَّم بيانٌ صحفي عن مشروع متصفِّح تُور قائمةً ببعض المواد الإخبارية التي تضمَّنت متصفِّح تُور خلال الفترة ما بين ١٩٩٩ و٢٠١٦.

أمَّا عن أوجه الاستخدام السلبية، فقد استخدمَ المجرمون الإلكترونيون شبكة تُور لإخفاء هُوياتهم على نطاق واسع. ويمكن الوصول إلى المواقع الإلكترونية عبر الخدمات التي جرى إخفاؤها بواسطة برنامج تُور، والتي تحتوي على اللاحقة الإنجليزية .onion. الكثير من هذه المواقع بغيضة للغاية، بما في ذلك المواقع غير القانونية على الويب المظلم، والتي تُستخدَم في تجارة المخدرات،

والإباحية، وغسل الأموال. على سبيل المثال، كان الوصول إلى موقع «سيلك روود»، وهو جزءٌ من الويب المظلم، ويشتهر بأنه منصة لبيع المخدرات وتوريد العقاقير المحظورة، يتم عبر متصفح تُور، ما صعَّب على جهاتٍ إنفاذ القانون تتبُّعه. بعد القبض على روس ويليام أولبريخت، كانت هناك محاكمة قضائية كبرى وأُدين بعد ذلك بتهمتَي إنشاء موقع «سيلك روود» وإدارته، تحت الاسم المستعار «القبطان الرهيب روبرتس». أغلق الموقع ولكنه عاود الظهور من جديد، وفي ٢٠١٦ ظهرت نسخته الثالثة الجديدة تحت اسم «سيلك روود ٣٫٠».

الويب الخفي

يشير الويب الخفي أو العميق (ديب ويب) إلى جميع المواقع التي لا يمكن فهرستها بواسطة محركات البحث المعتادة مثل جوجل، وبينج، وياهو. ويتضمَّن مواقع مشروعة بالإضافة إلى المواقع التي يتكوَّن منها الويب المظلم (دارك ويب). وتشير التقديرات إلى أنِ الويب العميق أكبر بكثير من الويب السطحي المألوف، ولكن يظل من الصعب تقدير حجم هذا العالَم الخفي من البيانات الضخمة حتى باستخدام محركات بحث مخصَّصة للويب الخفي.

الفصل الثامن

البيانات الضخمة والمجتمع

الروبوتات والوظائف

تنبَّأت كتاباتُ عالِمِ الاقتصاد البارز جون مينارد كينز خلال الكساد الاقتصادي البريطاني في ١٩٣٠ بما ستبدو عليه الحياةُ المهنية بعد قرنٍ من الزمن. خلقت الثورة الصناعية وظائفَ جديدةً في المصانعِ محورها المُدُن، وغيَّرت المجتمع الذي كان زراعيًّا في الأساس. كان يُعتقَد أن الأعمال التي تتطلَّب عددًا كبيرًا من العمالة ستؤديها الآلات في نهاية المطاف، الأمر الذي سيؤدي بالبعض إلى البطالة، وبالبعض الآخر إلى العمل لعدد قليل جدًّا من أيام الأسبوع. كان كينز مهتمًّا بوجه خاص بكيفية استخدام الناس لأوقات الفراغ الأطول بعد أن تُحرِّرهم التطورات التقنية من قيود المتطلبات المُلحَّة للعمل مقابل أجر. ربما كانت المسألة الأكثر إلحاحًا هي مسألة الدعم المالي التي تؤدي إلى الاقتراح بأن دخلًا أساسيًّا شاملًا من شأنه أن يوفِّر وسيلةً لمواكبة انخفاض عدد الوظائف المتاحة.

شهدنا تدريجيًّا، على مدار القرن العشرين، تناقُص عدد الوظائف في مجال الصناعة بسبب الآلات الأكثر تطوُّرًا، وعلى الرغم من أن الكثير من خطوط الإنتاج، على سبيل المثال، قد أصبحت آليةً بالكامل منذ عقود، فإن أسبوع العمل الذي يستمر لخمس عشرة ساعةً فقط الذي تنبَّأ به كينز لم يتحقَّق، ويبدو أنه كان من المُستبعَد أن يتحقَّق في المستقبل القريب. لا شك في أن الثورة الرقمية ستُغيِّر من أنماط العمالة، مثلما فعلت الثورة الصناعية تمامًا، ولكن بطرقٍ من المُستبعَد أن نتمكَّن من التنبؤ بها بدقة. ومع تطوُّر تقنية «إنترنت الأشياء»، أصبح اعتماد عالمنا على البيانات في تزايُد. سيلعب استخدام نتائج تحليل البيانات الضخمة في الوقت الحقيقي في اتخاذ القرارات والإجراءات دورًا تزداد أهميته في مجتمعنا يومًا بعد يوم.

ثمَّة مقترحات تقول إنَّ دور البشر سيقتصر فقط على صناعة الآلات وبرمجتها، ولكن هذا محض تخمين، كما أن هذا المجال، على أي حال، من مجالات العمل المتخصِّصة التي يمكننا أن نتوقَّع على نحو واقعي أن نرى الروبوتات تستبدل البشرَ فيها. على سبيل المثال، سيُقلِّل التشخيص الطبي الآلي المتطوِّر من عدد العِمالة الطبية. ومن المرجَّح أن يفعل الجرَّاحون الآليون، ذوو القدرات الكبيرة الشبيهة بقدرات نظام واتسون من شركة آي بي إم، المِثل. كما ستتطوَّر معالجة اللغات الطبيعية، وهي مجالٌ آخر من مجالات البيانات الضخمة، بدرجةٍ لن نتمكَّن معها من تمييز ما إذا كُنا نتحدَّث إلى جهاز آلي أم إلى طبيب، على الأقل عندما لا نتحدَّث إليه وجهًا لوجه.

ولكن، من الصعب التنبؤ بالوظائف التي سيؤديها البشر في حال سيطرت الروبوتات على الكثير من الأدوار الحالية. من المفترض أن يكون الابتكار مجالًا يخصُّ البشر دون غيرهم، إلا أنَّ علماء في مجال الكمبيوتر، يعملون بالتعاون فيما بينهم في جامعتَي كامبريدج وآبريستويث، طوَّروا عالِمًا آليًّا

أسمَوه آدم. نجحَ آدم في وضع فرضياتٍ جديدة في مجال علم الجينوم واختبارها، الأمر الذي أدَّى إلى اكتشافاتٍ علمية جديدة. وشهدت الأبحاث في هذا المجال تقدُّمًا أكبر عندما نجحَ فريقٌ من جامعة مانشستر في تطوير إيف، وهو روبوت يعمل على تصميم عقاقير للأمراض الاستوائية. وطبَّق كلا المشروعَين أساليبَ الذكاء الاصطناعي.

تتجلَّى براعة الروائيين على أنها ذات طابع بشري فريد؛ فهي نتاجُ الخبرات والمشاعر والخيال، ولكن حتى هذا المجال الإبداعي لم يسلم من غزو الروبوتات. تقبل جائزة نيكي هوشي شينيشي الأدبية رواياتٍ ألَّفَها أو شارك في تأليفها مؤلِّفون غير بشريين. في ٢٠١٦، اجتازت أربعُ رواياتٍ اشترك في تأليفها مؤلِّفون من البشر وأجهزة الكمبيوتر المرحلة الأولى من المسابقة، من دون أن يعلم الحُكَّام شيئًا عن تفاصيل تأليفها.

على الرغم من أن العلماء والروائيين قد ينتهي بهم المطاف بمشاركة العمل مع الروبوتات، فبالنسبة إلى أغلبنا، سيتجلَّى تأثيرُ البيئة القائمة على البيانات الضخمة على نحو أوضح في أنشطتنا اليومية؛ وذلك من خلال الأجهزة الذكية.

المَركبات الذكية

في ٧ ديسمبر ٢٠١٦، أعلنت أمازون أنها نجحت في جعل طائرتها الأولى من دون طيَّار لتوصيل الطلبات التجارية، تشق طريقها مسترشِدةً بنظام تحديد المواقع العالمي (جي بي إس). تسلم صاحبُ الطلب، وهو رجلٌ يعيش في الريف بالقرب من كامبريدج في المملكة المتحدة، طردًا يزن ٤٫٧ أرطال. يستفيد حاليًّا من خدمة توصيل الطلبات باستخدام طائراتٍ من دون طيَّار عميلان فقط من عملاء خدمة أمازون برايم إير، وكلاهما يعيشان ضمن مساحة تبلغ ٥٫٢ أميال فقط من مركز التوزيع بالقرب من كامبريدج. ثمَّة مقطع فيديو يعرض هذه الرحلة الجوية، وقد أشرنا إليه في قسم «قراءات إضافية». يبدو أن هذه الخدمة قد تكون إشارة البدء بجمع البيانات الضخمة من أجل هذا المشروع.

شركة أمازون ليست الشركة الأولى التي تنجح في توصيل الطلبات التجارية باستخدام طائراتٍ من دون طيَّار. في نوفمبر ٢٠١٦، بدأت شركة فليرتي في استخدام هذه الخدمة في توصيل البيتزا في حدود منطقة على مسافاتٍ صغيرة من مقرها في نيوزيلندا، كما كان يوجد عدد من المشروعات المشابهة في أماكن أخرى. يبدو حاليًّا أن خدمات التوصيل باستخدام طائراتٍ من دون طيَّار ستزداد، خاصةً في الأماكن المنعزلة حيث يمكن إدارة مسائل الخصوصية. لا شك أن هجومًا إلكترونيًّا أو حتى عطلًا في الأنظمة الحاسوبية من شأنه أن يتسبَّب في فوضى عارمة: إذا تعطَّلت طائرة توصيل صغيرة من دون طيَّار، على سبيل المثال، فقد تتسبَّب في إصابة أو وفاة البشر أو الحيوانات، كما أنها قد تتسبَّب في إلحاق أضرار جسيمة بالممتلكات.

هذا ما حدثَ عندما تمَّت السيطرة عن بُعد على البرنامج الذي يتحكَم في سيارة تسير على الطريق بسرعة ٧٠ ميلًا في الساعة. في ٢٠١٥، قدَّم خبيران أمنيان، تشارلي ميلر وكريس فالاسيك، يعملان في مجلة «وايرد»، عرضًا على متطوِّع لإثبات أن «يوكونيكت» Uconnect، وهي لوحة معلوماتٍ حاسوبية تُستخدَم في توصيل السيارة بالإنترنت، يمكن اختراقها عن بُعد أثناء تحرُّك السيارة. كانت نتائج التقرير مقلقة؛ فقد تمكَّن المخترقان الخبيران من استخدام كمبيوتر محمول متصل بالإنترنت في التحكم في سيارةٍ طراز جيب شيروكي على مستوى التوجيه والمكابح ونظام نقل الحركة، ووظائف أخرى أقل أهميةً مثل مكيِّف الهواء والراديو. كانت السيارة الجيب تتحرَّك بسرعة ٧٠ ميلًا في الساعة في طريق عام مزدحم عندما تعطَّلت استجابة دواسة السرعة تمامًا، الأمر الذي أفزعَ السائق كثيرًا.

نتيجةً لهذا الاختبار، أصدرت شركة كرايسلر العاملة في مجال تصنيع السيارات تحذيرًا إلى ١,٤ مليون مالِك سيارة وأرسلت إليهم محركات أقراص «يو إس بي» تحتوي على تحديثات برامج لتثبيتها عبر منفذٍ في لوحة المعلومات. نجحَ هذا الهجوم بسبب ثغرة أمنية في شبكة الهواتف الذكية تمَّ إصلاحها بعد ذلك، ولكن، توضِّح هذه القصة ضرورة التعامل مع فكرة احتمالية حدوث هجماتٍ إلكترونية على المركبات الذكية قبل أن تُصبح هذه التقنية متداولةً بالكامل.

يبدو أن حُلول المركبات الذاتية القيادة، بدءًا من السيارات إلى الطائرات، أمرٌ حتمي. أصبحت الطائرات تطير ذاتيًّا بالفعل، بما في ذلك الإقلاع والهبوط. وعلى الرغم من أن فكرة استخدام طائراتٍ من دون طيَّار في نقل البشر على نطاق واسع مُستبعَدة، فإنها تُستخدم حاليًّا في الزراعة في عملية الرش الذكي للمحاصيل، وكذلك في الأغراض العسكرية. ربما لا تزال المركبات الذكية في مراحل تطوُّرها الأولى لاستخدامها في الأغراض العامة، ولكن، أصبحت الأجهزة الذكية بالفعل جزءًا من المنازل الحديثة.

المنازل الذكية

كما ذكرنا في الفصل الثالث، يُعَد مصطلح «إنترنت الأشياء» طريقةً ملائمة للإشارة إلى الأعداد الهائلة من أجهزة الاستشعار الإلكترونية المتصلة بالإنترنت. على سبيل المثال، يُعَد أيُّ جهاز إلكتروني يمكن تركيبه في المنزل والتحكم فيه عن بُعد، من خلال واجهة مستخدِم يستعرضها قاطنُ المنزل عبر التلفزيون أو الهاتف الذكي أو الكمبيوتر المحمول، جهازًا ذكيًّا؛ ومن ثمَّ يكون جزءًا من إنترنت الأشياء. تُثبَّت نقاط تحكم مركزية تعمل بالصوت في الكثير من المنازل، والتي تتحكَّم في الإنارة، والتدفئة، وأبواب المرائب، والكثير من الأجهزة المنزلية الأخرى. يعني الاتصال بالواي فاي (تشير إلى «دقة النقل اللاسلكي»)، أو القدرة على الاتصال بشبكات، على غرار الإنترنت، باستخدام موجات الراديو بدلًا من الأسلاك) أنه يمكنك أن تسأل مكبِّر الصوت الذكي (عن طريق أن تدعوه بالاسم الذي ستُطلقه عليه) عن حالة الطقس المحلي أو التقارير الإخبارية الوطنية.

تقدِّم هذه الأجهزة خدماتٍ تستند إلى السحابة الإلكترونية، وهي لا تخلو من العيوب فيما يتعلق بالخصوصية. طالما أنَّ الجهاز قيد التشغيل، فكلُّ ما تقول يُسجَّل ويُخزَّن في خادم بعيد. خلال تحقيق في جريمة قتل حدثت مؤخَّرًا، طلبت الشرطة في الولايات المتحدة من شركة أمازون أن تُفصح عن بياناتٍ من أحد أجهزة إيكو (الذي يعمل بالتحكُّم في الصوت ويتصل بخدمة مساعِد أليكسا الصوتي لتشغيل الموسيقى، والتزويد بالمعلومات، والتقارير الإخبارية، وما إلى ذلك) اعتقادًا منهم أنها ستساعدهم في تحقيقاتهم. لم توافق شركة أمازون على فعل ذلك في البداية إلا أنَّ المشتبه به أذِنَ لها بالإفصاح عن التسجيلات أملًا في أنها ستساعد في إثبات براءته.

سيؤدِّي المزيد من التطور، بناءً على الحوسبة السحابية، إلى أن تصبح الأجهزة الكهربائية مثل الغسالات، والثلاجات، وروبوتات التنظيف المنزلية جزءًا من المنزل الذكي، ويتم التحكم فيها عن بُعد عبر الهواتف الذكية، أو أجهزة الكمبيوتر المحمولة، أو مكبِّرات الصوت المنزلية. وبما أنه يتم التحكُّم في جميع هذه الأنظمة عبر الإنترنت، فمن المُحتمَل أن تكون عُرضةً للخطر من قِبَل المخترقين؛ ومن ثمَّ، فإن الأمن مجال مهم يستوجب البحث.

حتى لُعب الأطفال ليست مُحصَّنة. فقد تعرَّضت دمية ذكية، تُدعَى «صديقتي كايلا»، حازت لقبَ «أفضل لعبة مبتكرة لعام ٢٠١٤» من اتحاد لندن لصناعة الألعاب، للاختراق بعد ذلك. يمكن للطفل، من خلال جهاز غير مُؤمَّن يعمل بالبلوتوث مخفي داخل الدمية، أن يطرح أسئلةً على الدمية ويسمع إجاباتها. حثَّت الوكالة الاتحادية للشبكات في ألمانيا، المسئولة عن مراقبة الاتصالات عبر الإنترنت، الآباءَ على تدمير الدمية، والتي مُنِع إنتاجها حاليًّا، بسبب ما تُمثِّله من خطر على الخصوصية. تمكَّن المخترقون من إثبات أنه من السهل للغاية أن يستمعوا إلى الطفل ويقدِّموا له إجاباتٍ غير مناسبة، بما في ذلك كلمات من قائمة الكلمات المحظورة التي وضعتها الشركة المُصنِّعة.

المدن الذكية

على الرغم من أن المنازل الذكية بدأت في التحوُّل إلى واقع، فمن المتوقع أن يحوِّل إنترنت الأشياء — بالإضافة إلى الأساليب المتعدِّدة لتكنولوجيا المعلومات والاتصالات — المدنَ الذكية إلى واقع. بدأت الكثير من الدول، بما فيها الهند، وأيرلندا، والمملكة المتحدة، وكوريا الجنوبية، والصين، وسنغافورة، في تصميم مدنٍ ذكية بالفعل. تدور فكرة المدن الذكية حول تحقيق فاعلية أكبر في عالَم اليوم المزدحم، وفي ظل النمو المطرد للمدن. يسجِّل انتقال سكان الريف إلى المدن معدلات ارتفاع متزايدة. في ٢٠١٤، كان ٥٤ في المائة من السكان يعيشون في المدن، وبحلول ٢٠٥٠، تتوقَّع الأممُ المتحدة أن حوالي ٦٦ في المائة من سكان العالَم سيقطنون المدن.

تُدفَع تقنية المدن الذكية بالأفكار المنفصلة المتراكمة من التطبيقات السابقة لإنترنت الأشياء وأساليب إدارة البيانات الضخمة. على سبيل المثال، ستكون السيارات من دون سائق، والمتابعة الصحية عن بُعد، والمنازل الذكية، والعمل عن بُعد من سمات المدينة الذكية. ستعتمد هذه المدينة على إدارة

وتحليل البيانات الضخمة المُجمَّعَة من جميع أجهزة الاستشعار الهائلة العدد في المدينة. ومن ثمَّ، فإن البيانات الضخمة وإنترنت الأشياء معًا هما جوهر المدن الذكية.

أمَّا عن أوجه النفع التي تعود على المجتمع ككل، فلعلَّ نظام الطاقة الذكي أحدها. من شأن هذا النظام أن ينظِّم إضاءة الشوارع، ومراقبة المرور، بل ومتابعة جمع القُمامة. ويمكن تحقيق هذا كلِّه من خلال تركيب عدد هائل من بطاقات تحديد الهُوية بموجات الراديو وأجهزة استشعار لا سلكية في جميع أنحاء المدينة. ستُرسل هذه البطاقة، المكوَّنة من شريحة دقيقة وهوائي صغير، البيانات من الأجهزة المنفردة إلى موقع مركزي لتحليلها. على سبيل المثال، يمكن لإدارة المدينة أن تتابع الحالة المرورية عن طريق تركيب بطاقات تحديد الهُوية بموجات الراديو في السيارات، وكذلك كاميرات رقمية في الشوارع. وسيكون الأمان الشخصي المُحسَّن أحد الاعتبارات أيضًا؛ إذ يمكن على سبيل المثال وضع بطاقاتٍ مع الأطفال سرًّا ومتابعتهم عبر الهاتف المحمول لأحد الوالدَين أو كلَيهما. ستُنتِج أجهزة الاستشعار هذه كميةً هائلة من البيانات التي ستحتاج إلى متابعة وتحليل في الوقت الحقيقي عبر وحدة معالجة بياناتٍ مركزية. ويمكن استخدامها بعد ذلك في مجموعة متنوِّعة من الأغراض، بما في ذلك قياسُ معدل الانسياب المروري، وتحديد مواقع الاختناقات المرورية، واقتراح مساراتٍ بديلة. ولا شكَّ أن أمن البيانات يُشكِّل أهميةً قصوى في هذا الإطار؛ فأي عطل أو اختراق كبير للنظام سيؤثِّر سريعًا في ثقة المواطنين.

أُنشِئت منطقة الأعمال الدولية في سونجدو بكوريا الجنوبية خصوصًا لتكون مدينةً ذكية. ومن بين السمات الرئيسية لهذه المدينة أنها تحتوي على اتصال واسع النطاق بالإنترنت عبر الألياف الضوئية. وتُستخدَم هذه التقنية الحديثة لضمان سرعة الوصول إلى السمات المرغوبة للمدينة الذكية. كما أن المدن الذكية الجديدة مُصمَّمة للحد من الآثار البيئية السلبية، ما يجعلها نموذج المدن المستقبلية المستدامة. في حين أن الكثير من المدن الذكية، مثل سونجدو، صُمِّمَت وأُنشِئت خصوصًا لهذا الغرض، فإن المدن الحالية ستستلزم تحديث بنيتها التحتية تدريجيًّا.

في مايو ٢٠١٦، كشفت مبادرة النبض العالمي التابعة للأمم المتحدة، وهي مبادرة تهدف إلى الترويج لأبحاث البيانات الضخمة من أجل الصالح العالمي، النقابَ عن مسابقتها المفتوحة تحت عنوان «مسابقة الأفكار العظيمة لعام ٢٠١٦: المدن المستدامة» للدول العشر الأعضاء في رابطة دول جنوب شرق آسيا ودولة كوريا. بحلول موعد المسابقة النهائي في شهر يونيو، جرى استلام أكثر من ٢٥٠ مُقترَحًا وأُعلِنَ عن الفائزين في العديد من الفئات في شهر أغسطس ٢٠١٦. فازت دولة كوريا بالجائزة الكبرى على مقترحها لتحسين وسائل النقل والمواصلات العامة عن طريق تقليل فترات الانتظار استنادًا إلى المعلومات المستقاة من الجمهور حول صفوف الانتظار.

استشراف المستقبل

في هذا الكتاب، رأينا كيف مرَّت بيانات العلوم بتحوُّلاتٍ جذرية على مدار العقود القليلة الماضية بفضل مظاهر التقدُّم التقني التي تحقَّقت بابتكار الإنترنت والكَون الرقمي. في هذا الفصل الأخير،

استرقنا النظر على بعض الجوانب في حياتنا التي تلعب البيانات الضخمة دورًا مهمًّا في تشكيلها، سواءٌ في الحاضر أو المستقبل. وعلى الرغم من أنه لا يمكننا أن نأمل في أن نغطِّي جميع الجوانب التي تؤثِّر فيها البيانات الضخمة في هذه المقدمة القصيرة، فقد تناولنا بعضًا من التطبيقات المتنوِّعة التي تؤثِّر فينا بالفعل.

ستزداد البيانات التي يُنتجها العالَم أكثر فأكثر. ولا شكَّ في أن أساليب التعامُل مع كل هذه البيانات بفاعلية وبطريقة مجدية ستظل موضوع الأبحاث المُكثَّفة، لا سيَّما في مجال التحليل في الوقت الحقيقي. تشير ثورة البيانات الضخمة إلى بداية تغيير جذري في الطريقة التي يسير بها العالم، وكما هو الحال مع جميع مظاهر التقدُّم التقني، أصبح الأفراد، والعلماء، والحكومات؛ مجتمعين يتحمَّلون مسئوليةً أخلاقية لضمان استخدامها على النحو الصحيح. البيانات الضخمة قوة. وإمكاناتها للخير هائلة. وكيفية تجنُّب إساءة استخدامها أمرٌ متروك لنا.

جدول سعة التخزين بالبايت

المصطلح	معناه
بت	رقم ثنائي واحد: صفر أو واحد
بايت	٨بت
كيلوبايت	١٠٠٠بايت
ميجابايت	١٠٠٠كيلوبايت
جيجابايت	١٠٠٠ميجابايت
تيرابايت	١٠٠٠جيجابايت
بيتابايت	١٠٠٠تيرابايت
إكسابايت	١٠٠٠بيتابايت
زيتابايت	١٠٠٠إكسابايت
يوتابايت	١٠٠٠زيتابايت

جدول الشفرة القياسية الأمريكية لتبادل المعلومات للأحرف الإنجليزية الصغيرة

الحرف	النظام السداسي العشري	النظام الثنائي	النظام العشري
a	٦١	٠١١٠٠٠٠١	٩٧
b	٦٢	٠١١٠٠٠١٠	٩٨

الحرف	النظام السداسي العشري	النظام الثنائي	النظام العشري
c	٦٣	٠١١٠٠٠١١	٩٩
d	٦٤	٠١١٠٠١٠٠	١٠٠
e	٦٥	٠١١٠٠١٠١	١٠١
f	٦٦	٠١١٠٠١١٠	١٠٢
g	٦٧	٠١١٠٠١١١	١٠٣
h	٦٨	٠١١٠١٠٠٠	١٠٤
i	٦٩	٠١١٠١٠٠١	١٠٥
j	A6	٠١١٠١٠١٠	١٠٦
k	B6	٠١١٠١٠١١	١٠٧
l	C6	٠١١٠١١٠٠	١٠٨
m	D6	٠١١٠١١٠١	١٠٩
n	E6	٠١١٠١١١٠	١١٠
o	F6	٠١١٠١١١١	١١١
p	٧٠	٠١١١٠٠٠٠	١١٢
q	٧١	٠١١١٠٠٠١	١١٣
r	٧٢	٠١١١٠٠١٠	١١٤
s	٧٣	٠١١١٠٠١١	١١٥
t	٧٤	٠١١١٠١٠٠	١١٦
u	٧٥	٠١١١٠١٠١	١١٧
v	٧٦	٠١١١٠١١٠	١١٨
w	٧٧	٠١١١٠١١١	١١٩

الحرف	النظام السداسي العشري	النظام الثنائي	النظام العشري
x	٧٨	٠١١١١٠٠٠	١٢٠
y	٧٩	٠١١١١٠٠١	١٢١
z	A7	٠١١١١٠١٠	١٢٢
مسطرة المسافة	٢٠	٠٠١٠٠٠٠٠	٣٢

قراءات إضافية

الفصل الأول: انفجار البيانات

David J. Hand, *Information Generation: How Data Rule Our World* (Oneworld, 2007).

Jeffrey Quilter and Gary Urton (eds), *Narrative Threads: Accounting and Recounting in Andean Khipu* (University of Texas Press, 2002).

David Salsburg, *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century* (W.H. Freeman and Company, 2001).

Thucydides, *History of the Peloponnesian War*, ed. and intro. M. I. Finley, trans. Rex Warner (Penguin Classics, 1954).

الفصل الثاني: لماذا البيانات الضخمة مميَّزة؟

Joan Fisher Box, *R. A. Fisher: The Life of a Scientist* (Wiley, 1978).

David J. Hand, *Statistics: A Very Short Introduction* (Oxford University Press, 2008).

Viktor Mayer-Schnberger and Kenneth Cukier, *Big Data: A Revolution That Will Transform How We Live, Work, and Think* (Mariner Books, 2014).

الفصل الثالث: تخزين البيانات الضخمة

C. J. Date, *An Introduction to Database Systems* (8th edn; Pearson, 2003).

Guy Harrison, *Next Generation Databases: NoSQL and Big Data* (Springer, 2015).

الفصل الرابع: تحليلات البيانات الضخمة

Thomas S. Kuhn and Ian Hacking, *The Structure of Scientific Revolutions: 50th Anniversary Edition* (University of Chicago Press, 2012).

Bernard Marr, *Big Data: Using SMART Big Data, Analytics and Metrics to Make Better Decisions and Improve Performance* (Wiley, 2015).

Lars Nielson and Noreen Burlingame, *A Simple Introduction to DataScience* (New Street Communications, 2012).

الفصل الخامس: البيانات الضخمة والطب

Dorothy H. Crawford, *Ebola: Profile of a Killer Virus* (Oxford University Press, 2016).

N. Generous, G. Fairchild, A. Deshpande, S. Y. Del Valle, and R. Priedhorsky, 'Global Disease Monitoring and Forecasting with Wikipedia', *PLoS Comput Biol* 10(11) (2014), e1003892. doi: 10.1371/journal.pcbi.1003892

Peter K. Ghavami, 'Clinical Intelligence: The Big Data Analytics Revolution in Healthcare. A Framework for Clinical and Business Intelligence' (PhD thesis, 2014).

D. Lazer and R. Kennedy, 'The Parable of Google Flu: Traps in Big Data Analysis', *Science 343* (2014), 1203–5. http://scholar.harvard.edu/files/gking/files/0314policyforumff.pdf.

Katherine Marconi and Harold Lehmann (eds), *Big Data and HealthAnalytics* (CRC Press, 2014).

Robin Wilson, Elizabeth zu Erbach-Schoenberg, Maximilian Albert, Daniel Power et al., 'Rapid and Near Real-Time Assessments of Population Displacement Using Mobile Phone Data Following Disasters: The 2015 Nepal Earthquake', *PLOS Currents Disasters*, Edition 1, 24 Feb 2016, Research Article. doi: 10.1371/currents.dis.d073fbece328e4c39087bc086d694b5c http://currents.plos.org/disasters/article/rapid-and-near-real-time-assessments-ofpopulation-displacement-using-mobile-phone-data-followingdisasters-the-2015-nepal-earthquake.

الفصل السادس: البيانات الضخمة والشركات الكبرى

Leo Computers Society, *LEO Remembered, By the People Who Worked on the World's First Business Computers* (Leo Computers Society, 2016).

James Marcus, *Amazonia* (The New Press, 2004).

Bernard Marr, *Big Data in Practice* (Wiley, 2016).

Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Harvard University Press, 2015).

Foster Provost and Tom Fawcett, *Data Science for Business* (O'Reilly, 2013).

الفصل السابع: أمن البيانات الضخمة وقضية سنودن

Andy Greenberg, *This Machine Kills Secrets* (PLUME, 2013).

Glenn Greenwald, *No Place to Hide: Edward Snowden, the NSA, and the U.S. Surveillance State* (Metropolitan Books, 2014).

Luke Harding, *The Snowden Files* (Vintage Books, 2014).

G. Linden, B. Smith, and J. York, 'Amazon.com Recommendations: Item-to-item Collaborative Filtering', *Internet Computing* 7(1) (2003), 76–80.

Fred Piper and Sean Murphy, *Cryptography: A Very Short Introduction* (Oxford University Press, 2002).

P. W. Singer and Allan Friedman, *Cyber security and Cyber war: What Everyone Needs to Know* (Oxford University Press, 2014).

Nicole Starosielski, *The Undersea Network* (Duke University Press, 2015).

Janet Vertesi, 'How Evasion Matters: Implications from Surfacing Data Tracking Online', *Interface: A Special Topics Journal* 1(1) (2015), Article 13. http://dx.doi.org/10.7710/2373-4914.1013, http://commons.pacificu.edu/cgi/viewcontent.cgi?article=1013&context=interface.

الفصل الثامن: البيانات الضخمة والمجتمع

Anno Bunnik and Anthony Cawley, *Big Data Challenges: Society, Security, Innovation and Ethics* (Palgrave Macmillan, 2016).

Samuel Greengard, *The Internet of Things* (MIT Press, 2015).

Robin Hanson, *The Age of Em* (Oxford University Press, 2016).

مواقع إليكترونية

https://www.infoq.com/articles/cap-twelve-years-later-how-the-rul eshave-changed.

https://www.emc.com/collateral/analyst-reports/idc-the-digitaluniv erse-in-2020.pdf.

http://newsroom.ucla.edu/releases/ucla-research-team-inventsnew 249693.

http://www.ascii-code.com.

http://www.tylervigen.com/spurious-correlations.

https://www.statista.com/topics/846/amazon.

https://www.wired.com/2015/07/jeep-hack-chrysler-recalls-1-4mve hicles-bug-fix.

http://www.unglobalpulse.org/about-new.

https://intelligence.house.gov/news.

http://www.unglobalpulse.org/about-new.

مصادر الصور

(2-1) A cluster diagram.

(2-2) Decision tree for transactions.

(3-1) Simplified view of part of a Hadoop DFS cluster.

(3-2) Graph database.

(3-3) A binary tree.

(3-4) The binary tree with a new node.

(3-5) Completed binary tree.

(4-1) Map function.

(4-2) Shuffle and reduce functions.

(4-3) Directed graph representing a small part of the Web.

(4-4) Directed graph representing a small part of the Web with added link.

مصادر الجداول

(2-1) Fraud dataset with known classifications.

(3-1) Key–value database.

(3-2) A coded character string.

(4-1) 10-bit array.

(4-2) Summary of hash function results.

(4-3) Bloom filter for malicious email addresses.

(4-4) Votes cast for each webpage.

(6-1) Books bought by Smith, Jones, and Brown.

(6-2) Jaccard index and distance.

(6-3) Star ratings for purchases.

Table of Contents